Activity choice modeling for pedestrian facilities

THÈSE N^{O} 6806 (2015)

PRÉSENTÉE LE 4 DÉCEMBRE 2015

À LA FACULTÉ DE L'ENVIRONNEMENT NATUREL, ARCHITECTURAL ET CONSTRUIT

LABORATOIRE TRANSPORT ET MOBILITÉ

PROGRAMME DOCTORAL EN GÉNIE CIVIL ET ENVIRONNEMENT

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Antonin DANALET

acceptée sur proposition du jury:

Prof. F. Golay, président du jury
Prof. M. Bierlaire, directeur de thèse
Prof. Y. Shiftan, rapporteur
Prof. F. C. Pereira, rapporteur
Prof. K. W. Axhausen, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2015

To Christina

Acknowledgements

This thesis has been a long and interesting journey that I did not embarked upon alone. This thesis contains the contributions of many colleagues and friends. Here are those who have accompanied me, supported me and encouraged me all along the writing of this dissertation.

First of all, I would like to express my deep gratitude to **Michel Bierlaire** for giving me the opportunity to perform this piece of research. He has helped me focus on the ideas contained in this thesis and has pushed me to reach a better and more formal version of these ideas. All his feedback, comments and thoughts have helped shape this thesis and my way of working. I would also like to thank him for the very pleasant working environment in the lab and the great non-academic time and discussions that we have had at conferences and elsewhere. After almost losing my job because of my failure to pronounce it properly, now I know how to say "Wittekop"[1].

I would also like to thank the members of my thesis jury, **Kay Axhausen**, **Francisco Pereira**, **Yoram Shiftan** and **François Golay**, for providing rich and constructive feedback on my dissertation, pertaining to the big picture as much as the technical details.

I am thankful towards all my colleagues at the Transport and mobility laboratory: **Flurin Hänseler** for correcting typos in this thesis, for help in writing the German abstract, as well as helpful discussions around the literature review, pedestrian congestion and nice discussions about Swiss politics and other topics; **Matthieu de Lapparent** for many discussions about methodological issues in Ch. 5 and about French politics; **Bilal Farooq** for his help with Ch. 3 and funny moments and discussions here in Switzerland; **Loïc Tinguely** for the implementation of the methodology in Ch. 5 and the generation of results. Special thanks go to **Jingmin Chen**, who helped me a lot for the methodological part of Ch. 3 and for the programming and data management during this first part of my research, and to **Gunnar Flötteröd**, who suggested strategic sampling for Ch. 4, shared his code and helped adapt it to my needs. I also thank all my colleagues in the lab, for lunch breaks, coffee breaks and Sat' breaks, both for professional and less professional discussions: **Marija Nikolic**, **Riccardo Scarinci**, **Aurélie Glerum**, **Thomas Robin**, **Anna Fernández Antolín**, **Eva Kazagli**, **Amanda Stathopoulos**, **Anne Curchod**, **Marianne Ruegg**, **Tomáš Robenek**, **Stefan Binder**, **Iliya Markov**, **Bilge Atasoy**, **Ricardo Hurtubia**, **Javi Cruz**, **Shadi Sharif Azadeh**, **Yousef Maknoun**, **Mila Bender**, **Jianghang Chen**, **Nitish Umang**, **Mamy Fetiarison**, **Silvia Varotto**, **Pascal Scheiben**, **Sohrab**

[1] A beer brewed in Belgium for the Swiss market...

Sahaleh, **Xinjun Lai**, **Olivier Gallay** and several others. You were a great team!

I would like to thank the people on EPFL campus who helped me and shared their knowledge: **Richard Timsit** collected the WiFi data and shared thought-provoking thoughts about research and the world with me, **Vincent Etter** answered my programming issues and shared interesting breaks in the ELA cafeteria, **Derek Christie** preciously helped me with the English. **Yves Bolognini** and **Florent Deseneux** provided data and their knowledge of the map of the campus. **Etienne Marclay** and **Nils Rinaldi** provided the point-of-sale data used in Ch. 5. I enjoyed discussing copyright and open access issues with **Raphaël Grolimund**, who helped me share my research data.

I had the good luck to have been playing a "home game" throughout my thesis, with my friends and family around me all the time. Indeed, my friends have been an important source of support and I would like to thank all of them for the great moments that we have spent together, **Xav**, **Oli**, **Baptistine**, **Luc**, **Mélanie**, **Camille**, **Léo**, **Tim**, my flatmates **Uli**, **Rosanna**, **Volo**, **Claude**, **Dom**, **Daniel** and **Alain**, and many others. I thank most warmly my parents, **Eliane** and **Pierre-André**, and also **Laurence** and **Thierry**, and my brother and sisters, **Coline**, **Sara** and **Renaud**, for being next to me, for their love and for their support.

Last but not least, I would like to tremendously and deeply thank **Christina** for being by my side. This thesis stole a bit of my time with her, but she fought back and we have spent amazing breathers, meals, week-ends and trips together. She has helped me reach the end of this thesis, but more than that, she is making my life better.

This research was supported by the Swiss National Science Foundation Grant 200021-141099 "Pedestrian dynamics: flows and behavior".

Lausanne, 26 October 2015 Antonin Danalet

Abstract

This thesis develops models of activity and destination choices in pedestrian facilities from WiFi traces. We adapt the activity-based travel demand analysis of urban mobility to pedestrians and to digital footprints. We are interested in understanding the sequence of activities and destinations of a pedestrian using discrete choice models and localization data from communication antennas.

Activity and destination choice models are needed by pedestrian facilities, in particular multimodal transport hubs such as train stations or airports, for decision aid when building new infrastructure, modifying existing infrastructures, or locating points of interest such as ticket machines in a train station. Understanding demand for activities is particularly important when facing an increasing number of visitors or when developing new activities, such as shopping or catering.

Data from existing sensors, such as WiFi access points, are cheap and cover entire facilities, but are imprecise and lack semantics to describe moving, stopping, destinations or activities carried out at destinations. Thus, understanding pedestrian behavior first requires to observe the actual behavior and detect stops at destinations, and second to model the behavior.

Part I of this thesis focuses on activity-episode sequence detection. We develop a Bayesian approach to merge raw localization data with other data sources in order to take into account the imprecision and describe activity-episode sequences. This approach generates several activity-episode sequences for a single individual. Each activity-episode sequence is associated with a probability of being the true sequence. It is based on a measurement equation and a prior probability distribution. The measurement equation expresses the imprecision of the sensor. The prior represents the attractivity of the different points of interest surrounding the measurement and allows the use of *a priori* information from other sources of data (register data, point-of-sale data, counting sensors, etc.).

Part II proposes models for activity and destination choices. The joint choice of activity type and activity timing is modeled with an activity path approach. The sequence of activity episodes is seen as a path in an activity network. Time is considered as discrete. Unlike traditional models, our model is not tour-based, starting and ending at the home location, since the daily “home”activity is meaningless in our context. The choice set contains all combinations of activity types and time intervals. The number of different paths is thus very large (increasing with time resolution and disaggregation of types of activities). Inspired by

route choice models, we use a Metropolis-Hastings algorithm for the sampling of paths to generate the choice set. An importance sampling correction of the utility allows the estimation of unbiased model parameters without enumerating the full choice set.

While the activity path model describes the choice of an activity type in time, the destination where the activity is performed is modeled with a destination choice model conditional on the activity type. Our approach accounts for the panel nature of the data and deals with serial correlation between error terms.

Using real WiFi data collected on the EPFL campus, we estimate an activity path choice model showing a satiation effect, a schedule delay effect related to class start time, primary activity preferences, time of day preferences and pattern preferences.

We also develop a destination model for a specific activity type: eating. Knowing that the individual has decided to eat, which restaurant does she choose? This conditional destination choice model includes in its utility the cost of menus, available types of foods and drinks, visibility of the restaurant, distance from a previous activity episode, socioeconomic characteristics and habits.

This thesis proposes a set of rigorous methodologies to detect, model and forecast pedestrian choices of activities and destinations in pedestrian infrastructure, using sensor data. A proof-of-concept has been developed on a campus using real data. Our decision-aid methodologies will help multimodal transport hub operators to optimize the locations of different points of interest (such as ticket machines, restrooms or shops), define opening hours or train schedules, and find a balance between different types of users (travelers or shoppers).

Key words: Activity choice; Destination choice; Network traces; Pedestrians; Semantically-enriched routing graph (SERG); Potential attractivity measure; Activity-episode sequence; Activity path; Activity network; Importance sampling; Strategic sampling; Dynamic model; Initial conditions problem; Panel data; Location choice

Résumé

Cette thèse développe des modèles de choix d'activité et de destination pour les infrastructures piétonnes à partir de traces WiFi. L'analyse de la demande en transport basée sur l'activité, initialement développée pour la mobilité urbaine, est ici adaptée aux piétons et aux empreintes numériques des téléphones portables. Nous nous intéressons à la compréhension des séquences d'épisodes d'activité d'un piéton en utilisant les modèles de choix discret et les données de localisation des antennes de communication.

Les modèles de choix d'activité et de destination sont nécessaires aux infrastructures piétonnes, en particulier pour les pôles d'échanges multimodaux comme les gares ou les aéroports, afin d'aider à la décision lors de la construction de nouvelles infrastructures, lors de la modification de structures existantes, ou pour le choix du positionnement de certains points d'intérêt tels que les distributeurs à billets dans les gares. Mieux comprendre la demande pour les différentes activités disponibles est particulièrement important face à l'augmentation du nombre de visiteurs ou lors du développement de nouvelles activités, telles que l'ouverture d'un magasin ou d'un lieu de restauration.

Les données issues de capteurs existants, tels que les antennes WiFi, sont bon marché et couvrent des infrastructures entières, mais elles sont imprécises et dépourvues de la sémantique permettant de décrire les mouvements, les arrêts, les destinations et les activités effectuées à destination. Dès lors, comprendre les comportements piétons nécessite tout d'abord d'observer le comportement réel et de détecter les arrêts aux destinations, et seulement ensuite de modéliser le comportement.

La première partie de cette thèse se focalise sur la détection des séquences d'épisodes d'activité. Nous y développons une approche bayésienne pour fusionner des données de localisation brutes avec d'autres sources de données de manière à prendre en compte l'imprécision de la localisation et à décrire les séquences d'épisodes d'activité. Cette approche génère plusieurs séquences d'épisodes pour un individu. Chacune d'entre elles est associée à la probabilité d'être la vraie séquence, effectivement effectuée par l'individu. L'approche s'appuie sur une équation de mesure et une distribution de probabilité *a priori*. L'équation de mesure exprime l'imprécision du capteur. La distribution *a priori* représente l'attractivité des différents points d'intérêt dans le voisinage de la mesure et permet l'utilisation d'information *a priori* à partir d'autres sources de données (données de registres, données de points de vente, capteurs de comptage, etc.).

Abstract

La deuxième partie de cette thèse propose des modèles pour les choix d'activité et de destination. Le choix commun du type d'activité et du timing des activités est modélisé avec une approche dite "du chemin d'activité". La séquence d'épisodes d'activité est vue comme un chemin dans un réseau d'activité. Le temps est considéré comme discret. Contrairement aux modèles traditionnels, notre modèle n'est pas basé sur la notion de *tour*, commençant et terminant au domicile, puisque le domicile comme activité quotidienne n'a pas de sens dans notre contexte. L'ensemble de choix contient toutes les combinaisons de types d'activité et d'intervalles de temps. Le nombre de chemins d'activité différents est par conséquent très élevé (augmentant avec la résolution temporelle et la désagrégation des types d'activité). Inspirés par les modèles de choix d'itinéraire, nous utilisons un algorithme de Metropolis-Hastings pour échantillonner des chemins, générant ainsi l'ensemble de choix. Une correction de l'utilité liée à l'échantillonnage préférentiel permet d'estimer des paramètres du modèle non biaisés sans énumérer l'ensemble de choix complet.

Alors que le modèle de chemin d'activité décrit le choix d'un type d'activité dans le temps, la destination où cette activité a lieu est modélisée à l'aide d'un modèle de choix de destination conditionnel au type d'activité. Notre approche prend en compte la nature de panel des données et gère l'autocorrélation entre les termes d'erreur.

En utilisant des données WiFi réelles collectées sur le campus de l'EPFL, nous estimons un modèle de choix de chemin d'activité qui montre un effet de satiété, un effet d'aversion au retard lié aux heures de cours, une préférence pour une activité principale, une préférence pour l'heure de la journée et une préférence pour certains profils d'activité (c'est-à-dire la préférence pour un ordre dans lequel réaliser les activités).

Nous développons aussi un modèle de choix de destination pour un type d'activité spécifique : manger. Sachant que l'individu a décidé de manger, quel restaurant choisit-il ? Ce modèle de choix de destination conditionnel inclut dans sa fonction d'utilité le coût des menus, le type de nourriture et de boisson proposées, la visibilité du restaurant, la distance à partir de l'épisode d'activité précédent, les caractéristiques socioéconomiques et les habitudes.

Cette thèse propose un ensemble de méthodologies rigoureuses pour détecter, modéliser et prédire les choix d'activité et de destination des piétons dans les infrastructures qui leur sont dédiées, en utilisant des données de capteurs. Une démonstration de faisabilité a été développée sur le campus en utilisant des données réelles. Nos méthodologies d'aide à la décision aideront les opérateurs des pôles d'échanges multimodaux à optimiser la localisation des différents points d'intérêt (tels que distributeurs à billets, toilettes ou magasins), à définir les heures d'ouverture ou les horaires de train, et à trouver un équilibre entre les différents types d'usagers (voyageurs ou personnes qui font leurs courses).

Mots clefs : Choix d'activité ; Choix de destination ; Traces WiFi ; Piétons ; Réseau de routage enrichi sémantiquement ; Mesure d'attractivité potentielle ; Séquence d'épisodes d'activité ; Chemin d'activité ; Réseau d'activité ; Echantillonnage préférentiel ; Modèle dynamique ; Problème des conditions initiales ; Données panel ; Choix de localisation

Zusammenfassung

Die vorliegende Arbeit befasst sich mit der Entwicklung von Modellen, mit welchen die Wahl von Destinationen und Aktivitäten in Fussgängeranlagen anhand von 'WiFi-Spuren' vorhergesagt werden kann. Die aktivitäten-basierte Mobilitätsanalyse, welche ursprünglich für die städtische Mobilität entwickelt wurde, wird auf Fussgänger und digitale 'Fussabdrücke' von Mobiltelefonen angewandt. Wir interessieren uns für das Verständnis von Aktivitäts- und Destinations- Sequenzen von Fussgängern und verwenden dafür diskrete Entscheidungsmodelle in Verbindung mit Lokalisierungsdaten von Kommunikationsantennen.

Diskrete Entscheidungsmodelle für die Wahl von Aktivitäten und Destinationen finden Anwendung in Fussgängerinfrastrukturen, insbesondere bei multimodalen Verkehrsknoten, wie Bahnhöfen oder Flughäfen, bei der Dimensionierung neuer Anlagen, bei Umbauten oder bei der Platzierung besonderer Dienstleistungsstellen, wie Billett-Automaten. Das Verständnis der Aktivitätsnachfrage ist besonders wichtig bei steigenden Besucherzahlen oder bei einer Angebotserweiterung, wie zum Beispiel der Eröffnung eines Ladens oder eines Imbiss-Standes.

Daten bestehender Sensoren wie WiFi-Antennen sind kostengünstig und decken oft die gesamte Infrastruktur ab. Gleichzeitig sind sie aber auch ungenau und es fehlt ihnen eine Semantik zur Beschreibung von Bewegungen, von Aufenthalten, von Destinationen sowie von Aktivitäten, denen an einzelnen Orten nachgegangen wird. Für das Verständnis des Fussgängerverhaltens ist es somit unumgänglich, zunächst das tatsächliche Verhalten zu beobachten und Aufenthalte an Destinationen zu erkennen. Erst anschliessend kann das eigentliche Verhalten modelliert werden.

Der erste Teil dieser Arbeit beschäftigt sich mit der Erkennung von Aktivitätsepisodensequenzen. Wir entwickeln einen Bayes'schen Ansatz für die Verknüpfung von Lokalisierungsrohdaten mit anderen Datenquellen, um die Ungenauigkeit der Daten zu berücksichtigen und um die Aktivitätsepisodensequenzen zu beschreiben. Dieser Ansatz generiert mehrere Aktivitätsepisodensequenzen für jeden Fussgänger. Jede von ihnen hat eine gewisse Wahrscheinlichkeit, die wahre Sequenz zu sein. Das Verfahren stützt sich auf Messgleichungen und Wahrscheinlichkeitsverteilungen, welche vorgängig bekannt sind. Die Messgleichung drückt die Ungenauigkeit des Sensors aus. Die A-priori-Wahrscheinlichkeit stellt die Attraktivität von verschiedenen Destinationen in der Umgebung des Messortes dar und erlaubt es, vorgängig bekannte Information von anderen Datenquellen (Registrierungsdaten, Verkaufsdaten, Zählsensoren, etc.) zu berücksichtigen.

Abstract

Im zweiten Teil der Arbeit werden Modelle zur Wahl von Aktivitäten und Destinationen präsentiert. Die gleichzeitige Wahl des Aktivitätstypes und der Aktivitätszeit wird mit einem sogenannten “Aktivitätspfadansatz” modelliert. Die Aktivitätsepisodensequenz wird dabei als ein Pfad in einem Aktivitätsnetzwerk angesehen, wobei die Zeit als diskret betrachtet wird. Im Gegensatz zu klassischen Modellen basiert das unsere nicht auf einer Tour, die zu Hause beginnt und endet, da die Aktivität „Zuhause“ in unserem Kontext keine eigentliche Bedeutung hat. Die Auswahlmöglichkeiten beinhalten alle Kombinationen von Aktivitätstypen und Zeitintervallen. Die Anzahl von unterschiedlichen Aktivitätspfaden ist demzufolge sehr gross (sie nimmt mit der zeitlichen Auflösung und der Feinheit der Unterscheidung an Aktivitätstypen zu). Inspiriert von Routenwahlmodellen verwenden wir ein Metropolis-Hastings-Algorithmus, um eine Auswahl an Pfaden zu genieren. Eine Varianzreduktionstechnik erlaubt es, erwartungstreue Modellparameter zu schätzen ohne die Gesamtheit aller Pfade zu berücksichtigen. Während das Aktivitätspfadmodell die Wahl eines Aktivitätstypes und eines Zeitintervalls beschreibt, wird die Wahl der Destination, an welcher die Aktivität ausgeführt wird, mit einem auf den Aktivitätstyp bedingten Destinationswahlmodell beschrieben. Unser Ansatz berücksichtigt Paneleffekte der Daten, sowie Reihenkorrelation zwischen Fehlertermen.

Basierend auf realen WiFi-Daten vom Universitätscampus der ETH Lausanne schätzen wir ein Aktivitätspfadwahlmodell, welches Effekte wie die Sättigung, die Vermeidung von Verspätungen bezüglich der Vorlesungszeiten sowie Vorlieben für Hauptaktivitäten, für Tageszeiten und für verschiedene Verhaltensmuster darstellt.

Wir entwickeln ebenfalls ein Destinationswahlmodell für den spezifischen Aktivitätstyp ‘Essen’. Angenommen, eine Person entscheidet sich etwas zu essen, welchen Verpflegungsort (Restaurant, Kantine, Take-Away) wird sie wählen? Dieses bedingte Destinationswahlmodell berücksichtigt in seiner Nutzenfunktion die Verpflegungskosten, die Art der verfügbaren Ess- und Trinkmöglichkeiten, die Sichtbarkeit eines Verpflegungsortes, die Entfernung zur vorhergehenden Aktivitätsepisode, sowie sozioökonomische Eigenschaften und Angewohnheiten.

Diese Arbeit präsentiert eine Reihe rigoroser Methodologien zur Erkennung, Modellierung und Vorhersagung von Aktivitäten und Destinationen in Fussgängeranlagen basierend auf Sensordaten. Eine Machbarkeitsstudie mit realen Daten auf einem Universitätscampus wurde durchgeführt. Unsere Methoden werden den Betreiber von multimodalen Verkehrsknotenpunkten helfen, die Platzierung von Dienstleistungsorten wie Billett-Automaten, Toiletten oder Verkaufsstellen zu optimieren, die Öffnungszeiten oder den Zugfahrplan zu bestimmen, und ein Gleichgewicht zwischen Kundentypen (Reisende oder Personen, welche ihre Einkäufe tätigen) zu finden.

Stichwörter: Aktivitätswahl; Destinationswahl; WiFi-Spuren; Fussgänger; semantisch angereicherte Routing-Graphen; potentielles Attraktivitätsmass; Aktivitätsepisondesequenz; Aktivitiätspfad; Aktivitätsnetzwerk; Varianzreduktion; dynamisches Modell; Anfangswertproblem; Paneldaten; Standortwahl

Contents

List of Figures

List of Tables

List of symbols

Note that there is a conflict in notation between Ch. 3 and Ch. 5. In Ch. 3, the location of the activity episode $a = (x, t^-, t^+)$ is represented by x, $x \in POI$. In Ch. 5, the location of the activity episode is represented by i; it is the typical notation in the choice modeling community. The notation x has been kept in Ch. 3, together with $\hat{x}$, the position of the raw measurement $\hat{m}$, and $\mathring{x}$, the actual location of measurement location $\hat{x}$.

a_{ψ_n}	an activity episode, $a_{\psi_n} = (x, t^-, t^+)$, for individual n. Indexed by ψ
$a_{1:\Psi_n} = (a_1, ..., a_\psi, ..., a_{\Psi_n})$	an observed activity-episode sequence for individual n, abbreviated $a_{1:\Psi}$
$att_n(x,t)$	the attractivity for location $x \in POI$ at time t, for individual n
A_ψ	an activity, $A_\psi = (\mathscr{A}_k, t^-, t^+)$. Indexed by ψ
$A_{1:\Psi_n} = (A_1, ..., A_\psi, ..., A_{\Psi_n})$	an activity pattern with Ψ_n activities, indexed by ψ
$A(a_{\psi_i})$	a function $a_{\psi_i} \mapsto A(a_{\psi_i}) = \mathscr{A}_k \in \{\mathscr{A}_1, \mathscr{A}_2, ..., \mathscr{A}_K\}$
$\mathscr{A}_k$	an activity type, $k \in K$
$\mathscr{A}_{k,\tau}$	a node in the activity network, corresponding to activity type $\mathscr{A}_k$ and unit of time τ
$\mathscr{A}_{1:T}$	an activity path, i.e., a representation of an activity pattern $A_{1:\Psi_n}$ in an activity network
α	the parameter associated with travel time in the schedule delay approach
α_{in}	the agent effect for individual n and location i (time-invariant; "between" individuals variability)
α_n	the set of agent effects α_{in} for individual n and all destinations i, $\alpha_n = \{\alpha_{in}, \forall i\}$, $\alpha_n \in \mathbb{R}^{dim(i)}$
b	the parameter associated with the first choice y_{in0} in the model of the agent effect α_{in}
$b(\Gamma)$	the unnormalized target weights for path Γ
β	the parameters of the choice model
c	the parameter of the vector of time-invariant explanatory variables $\bar{x}_n$ in the model of the agent effect α_{in}
$\mathscr{C}_n$	the choice set for an individual n
$\mathscr{C}_{nt}$	the choice set for an individual n at time t

d	distance between the iterations in the Metropolis-Hastings algorithm
DDR	the domain of data relevance
δ	a cost function
$\delta_v(v)$	the cost of node v
$\delta_\Gamma(\Gamma)$	the non-link-additive cost of path Γ
$\delta(\Gamma)$	the generalized cost of path Γ
e	the end node of an activity network
ε_{int}	the error term for location i, individual n and time t. We assume $\varepsilon'_{int} \sim EV(0,1)$
ε'_{int}	the error term for location i, individual n and time t (short-term variation of preferences of individual n; "within" an individual variability). We assume $\varepsilon'_{int} \sim EV(0,1)$
$\mathscr{E}$	the set of edges in $SERG$
η_k	the parameters for satiation for activity type $\mathscr{A}_k$
f	the labeling function, $f : \mathscr{N} \to \mathscr{L}$, in $SERG$
g	a function associating nodes with coordinates in a coordinate system, $g : \mathscr{N} \to \mathbb{R} \times \mathbb{R} \times \mathbb{R}$
γ_e, γ_l	the parameters for early and late schedule delay
Γ	a path in the activity network
i	location of an activity episode a_{ψ_n} in the choice context of Ch. 5 (corresponds to x in Ch. 3)
j	the index of a measurement $\hat{m}_j$
J_n	the total number of measurements for individual n (abbreviated J)
k	the index of an activity type.
$k_{\Gamma n}$	the number of times activity path Γ is drawn
K	the number of activity types.
L	the number of different activity-episode sequences $a_{1:\Psi_n}$ corresponding to individual n
$\mathscr{L}$	a set of relevant labels for rooms in $SERG$
$\hat{m}_j$	a raw measurement, containing location $\hat{x}$ and timestamp $\hat{t}$. Indexed by j
$\hat{m}_{1:J_n,n}$	a set of J_n measurements $\hat{m}_j$ for individual n
μ	scale parameter for the Metropolis-Hastings algorithm, $\mu \geqslant 0$, $\mu = \frac{\ln 2}{(\zeta-1)\delta_{SP}}$
n	an individual
$\mathscr{N}$	the set of all nodes in $SERG$
POI	the set of points of interest, $POI \in \mathscr{N}$
$\mathscr{P}_i$	the set of all candidate activity paths for observation i
$\mathscr{P}_I$	the set of all candidate activity paths for all individuals $i \in I$ in the period of interest.

$\mathscr{P}_{A_{1:\Psi_i}}$	the set of candidate paths corresponding to the activity pattern $A_{1:\Psi_i}$
ψ_n	the index of a activity episode a_{ψ_n}.
Ψ_n	the total number of episodes a_{ψ_n} in the activity-episode sequence $a_{1:\Psi_n}$ and the total number of activities A in the activity pattern $A_{1:\Psi_n}$. Ψ_n is individual specific (abbreviated Ψ)
$q(j)$	the sampling probability
ρ	the parameter of the lagged variable $y_{xi(t-1)}$.
s	the start node of an activity network
$S_{x,i}(t)$	the instantaneous potential attractivity measure in location $x \in POI$ at time t for individual i
$S_{x,i}(t^-, t^+)$	the potential attractivity measure in location $x \in POI$ between start time t^- and end time t^+ for individual i
$sched_{x,i}(t)$	a dummy variable for time constraints in location $x \in POI$ at time t for individual i
SDE, SDL	early and late schedule delay, defined as $SDE = \max(t^* - t^a, 0)$ and $SDL = \max(t^a - t^*, 0)$
$SERG$	a semantically-enriched routing graph, $SERG := (\mathscr{N}, \mathscr{E}, \mathscr{L}, f, g, POI)$
σ_{α_i}	the parameters of the normal distribution in the model of the agent effect α_{in}. The i subscript is omitted to make the notation light
Σ_{α_i}	the matrix of parameters for the normally distributed error term ξ_{in} in the model of the agent effect α_{in}, $\xi_{in} \sim N(0, \Sigma_\alpha)$, $\Sigma_{\alpha_i} = \sigma^2_{\alpha_i} I$. The i subscript is omitted to make the notation light
t	time
τ	a discrete unit of time in the activity network, $\tau \in 1, 2, ..., T$. τ can also be seen as a time interval between τ_{LB} and τ_{UB}
τ_{LB}	the lower bound of the time interval τ
τ_{UB}	the upper bound of the time interval τ
T	the total number of units of time τ (in Ch. 4) or the total number of observations in panel data (in Ch. 5)
T_{min}	a minimum time threshold for activity episodes (typically 5 minutes in a pedestrian context)
$\hat{t}$	a timestamp of a raw measurement $\hat{m}$, continuous
t^a	the actual arrival time in the schedule delay approach
t^-	the start time of an activity episode a, a continuous random variable
t^+	the end time of an activity episode a, a continuous random variable

t^*	the preferred arrival time in the schedule delay approach
tt	the travel time to destination in the schedule delay approach
$tt_{x_\psi, x_{\psi+1}}$	the travel time from x_ψ to $x_{\psi+1}$.
U_{int}	the utility of location i for individual n at time t.
$\mathscr{U}$	the choice set corresponding to the activity network
v	a node in the activity path Γ, $v \in \Gamma$
$w(\Gamma)$	the target weight of Γ
x	the episode location, $x \in POI$
$\hat{x}$	the position of a raw measurement $\hat{m}$. $\hat{x} \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}$ (x-y coordinates in a coordinate system, and floor or altitude in a multi-floor environment).
$\bar{x}_n$	the vector of time-invariant explanatory variables, i.e., long-term preferences, socioeconomic characteristics
$\mathring{x}$	actual location of measurement location $\hat{x}$
ξ	accuracy of a measurement location $\hat{x}$
ξ_{in}	the normally distributed error term in the model of the agent effect α_{in}, $\xi_{in} \sim N(0, \Sigma_\alpha)$
y_{int}	choice of location i at time t with activity type k by individual n.
$y_{in(t-1)}$	lagged variable, represents the choice of location i for the previous activity episode with the same activity type k performed by the same individual n as compared to y_{int}
y_{in0}	the first choice in the observed panel data, also called initial value
y_{int}^{count}	the count of previous choices of alternative i by individual n up to the time t of the current choice, $y_{int}^{\text{count}} = \sum_{t'=0}^{t-1} I(yint')$

1 Introduction

This thesis seeks to develop a better understanding of pedestrian demand and to develop decision aid tools. It builds an activity-based approach from communication network traces, modeling activity and location choices. This approach is sensitive to policies and demand management strategies, when modifying or building pedestrian infrastructures or designing public transit timetables. Network traces, such as WiFi signatures, allow the observation and modeling of behavior.

1.1 Thesis motivation: Pedestrian activities in public spaces

1.1.1 Understanding pedestrian demand

Crowd dynamics and pedestrian modeling have been extensively studied in recent years due to urban growth and its pressure on urban infrastructure (Bierlaire and Robin; 2009; Duives et al.; 2013; Kasemsuppakorn and Karimi; 2013; Kneidl et al.; 2013; Weidmann et al.; 2014) and due to the availability of new tracking data (Sevtsuk et al.; 2009; Naini et al.; 2011; Versichele et al.; 2012; Duives et al.; 2014; Yoshimura et al.; 2014; van den Heuvel et al.; 2015). Crowd and pedestrian simulation is emerging as a tool for designing new infrastructures and optimizing the use of current infrastructures (Tabak et al.; 2010; Kim et al.; 2015).

Understanding pedestrian demand is important for several reasons. In shops, the analysis of the factors impacting purchase behavior are useful for marketing purposes (e.g., Hui et al.; 2009; Kholod et al.; 2010; Yaeli et al.; 2014). Visitor counting in different parts of a building is useful for demand-controlled ventilation, improving indoor air quality and energy savings (e.g., Tabak et al.; 2010; Kuutti et al.; 2014). For tourism, the visitor's experience can be improved based on people's behavior, through way-finding systems in city centers (Kemperman et al.; 2009; Edwards and Griffin; 2013), management of congestion in museums (Yoshimura et al.; 2014), or development of infrastructure in parks (O'Connor et al.; 2005). Mass events such as festivals can evaluate the number of visitors (Naini et al.; 2011) and their daily and hourly patterns of visits (Versichele et al.; 2012). In hospitals, Lee et al. (2012) optimize the

allocation of facilities on a floor and Haq and Luo (2012) review studies on nurses' behavior, patient preferences, building development and extension design. In inner-city shopping areas or Central Business Districts, pedestrian demand depends on the properties and spatial distribution of facilities, and conversely the economic viability of facilities depends on pedestrian behavior (Borgers and Timmermans; 1986b; Saarloos et al.; 2010).

1.1.2 Decision aid tools for public spaces

Operators of pedestrian facilities are interested in pedestrian demand, i.e., in (1) knowing how many people are visiting, (2) what visitors are doing once in the facility and (3) for how long. In facilities with controlled entrances, such as museums, counting can be easily performed with mechanical doors or sales data. When it is not the case, typically in transport hubs, counting is already a challenging task. Determining the activities of visitors and their duration in the facility is more difficult. What are shoppers buying in shops? What are visitors observing in museums? Are people in transport hubs shoppers or travelers? And do travelers need to buy a ticket or not?

Issues such as the source of congestion or the location of points of interest in pedestrian facilities need a solution. A decision aid tool for public spaces must predict the total activity travel demand within a public space. This requires the development of demand models at the scale of pedestrian infrastructures. It will help shops in defining their layout, music festival in locating toilets or optimizing concert schedules, transport hubs in locating ticket machines, or museums in setting up exhibitions avoiding bottlenecks.

1.2 Thesis objective: An activity-based approach to pedestrian demand

Pedestrian demand is driven by a need to perform activities in different locations. The existence of time-space constraints in pedestrian infrastructures asks for explicit modeling of activity scheduling decisions. Such models are traditionally used for people performing trips at an urban scale as an important source of information for strategic planning, and management or optimization of transportation networks (Ben-Akiva et al.; 1996; Bowman and Ben-Akiva; 2001; Arentze and Timmermans; 2004; Balmer et al.; 2006; Roorda et al.; 2008, among others). For pedestrians, they are useful in describing congestion, for the efficient design of new facilities, and for travel guidance and information systems.

The goal of this thesis is to adapt the activity-based approaches developed since the 1970s for urban areas to pedestrian facilities. Such models are sensitive to demand management strategies such as modifying schedules (e.g., train schedules in train stations or concert schedules in a music festival) and to land-use policies (e.g., concentrating the development of services in the center of a facility or in corridors). Activity-based modeling can also evaluate the impact of information and communication technology in public spaces, like in airports (Kalakou et al.;

2015) or in train stations, e.g., on the purchase of the ticket (going to the ticket machine *versus* buying on the phone).

1.2.1 Pedestrian choice of activity patterns

Published pedestrian research focuses mostly on walking behavior and crowd dynamics, as proven by the many reviews on various walking behavior models (Papadimitriou et al.; 2009; Schadschneider et al.; 2009; Bellomo et al.; 2012; Duives et al.; 2013; Mustafa and Ashaari; 2015). Bierlaire and Robin (2009) decompose pedestrian behavior into activity, destination, mode, route, next step and speed choices. From their review of the literature, it appears that the least studied area is activity choice for pedestrians. Borgers and Timmermans (1986b) sequentially model the choice of a destination in time. Then, they consider impulse shopping stops on the way, conditional on destination and route choice. Hoogendoorn and Bovy (2004) decompose pedestrian behavior into three levels: strategic (departure time choice, activity pattern choice), tactical (activity scheduling, destination choice, route choice) and operational (walking behavior). They focus on the simultaneous choice of route and destination and minimize a cost function depending on a walking cost and an activity scheduling cost. The activity scheduling cost is used to control the order of activities, the "mandatory" activities and the schedule delay. The cost function must be evaluated and compared to all other orders of activities. All possible orders of activities quickly becomes very large, when the number of activities increases. They propose an example with two activities, 1 and 2, leading to two possible orders of activities, 12 or 21. Daamen (2004, ch. 2) mentions activity patterns modeling as a blank spot in research. Liu, Usher and Strawderman (2014) model the activity pattern as the choice of performing an activity type in one of three time intervals delimited by check-in and security control in an airport. This approach does not model duration nor time-of-day preferences.

For car trips and at urban scale, activity-based models often assume a "home" location and a tour-based structure (Ben-Akiva et al.; 1996; Shiftan; 1998; Bhat and Singh; 2000; Bowman and Ben-Akiva; 2001; Miller et al.; 2005; Shiftan; 2008; Abou-Zeid and Ben-Akiva; 2012), which is not very well adapted for pedestrian facilities. The choices of activity types and scheduling are often not considered simultaneously, for combinatorial reasons (Shiftan and Ben-Akiva; 2011). When considering simultaneously activity type and scheduling, assuming that the period of interest contains T time units, the number of alternatives is K^T and the choice set size increases exponentially. By considering tours or sequential order, without timing, the size of the choice set decreases. Other approaches consider time as a continuous variable and use a multiple discrete-continuous approach (Bhat; 2005) or a dynamic discrete-continuous approach (Habib; 2011).

A brief review of activity choice models for pedestrians and for urban areas shows that there is no simultaneous model for a full sequence of activities, or activity pattern, including preferred time of day, preferred arrival time and preferred duration. In the general literature

about activity-choice, the combinatorial number of activity sequences is solved by modeling subproblems (e.g., activity patterns without time of day and duration) and by assuming *a priori* structures in activity patterns (e.g., mandatory vs discretionary). Models are tour-based, assuming a *home* location. Therefore, they cannot be directly applied to pedestrian facilities. This research gap needs to be filled with a model of activity patterns adapted to pedestrian facilities, without assumptions about priorities of activities and including preferences that define order of activities, time of day and duration simultaneously.

1.2.2 From geolocalized data to activity and destination choices

Our objective is to develop a complete activity-based pedestrian demand model. The traditional activity-schedule approach decomposes the activity-travel decision into two sets of models (Abou-Zeid and Ben-Akiva; 2012): a *daily activity pattern model* including the number and purposes of tours and the number of stops per tour, and *tour-level models* including the destination and mode choice and the timing of activities. Our goal is to merge the daily activity pattern models with the timing of activities from tour-level models from the traditional activity-schedule approach. We assume that all visitors walk, therefore we do not consider mode choice in the context of pedestrian facilities, and so we do not consider it in this thesis.

A complete activity-based pedestrian demand analysis requires real data for the estimation and evaluation of the methods. Our objective consists in using existing data: network traces, measures of attractivity and map information. Individual mobility traces are becoming available from pervasive systems, such as cellular networks (González et al.; 2008) or WiFi hotspots (Section 2.1.1). In many cases, cost and privacy issues do not allow the use of high precision sensors such as cameras covering an entire pedestrian infrastructure. The large size of certain public spaces, such as an airport or a music festival, implies either precise sensors with incomplete coverage (e.g., cameras or bluetooth sensors in intersections), or full coverage with imprecise long range sensors (e.g., cellular network data, traces from WiFi infrastructures). Apart from being imprecise, network traces follow individuals over a longer period than traditional pen-and-paper surveys (see Section 2.2.1). Thus, it becomes possible to collect activity-episode sequences covering several days, weeks or months. Network traces are not the only available data to help understand activity and location choices. Aggregate measures of attractivity, i.e., occupancy, are also available, such as point-of-sale data or occupancy of a train or airplane, associated with a schedule. Map information, such as distances from a pedestrian network or the location of points of interest, help the understanding of pedestrian behavior. In this thesis, we develop methodologies to merge these different sources of data and use their panel nature.

1.3 Thesis contributions: Activity episodes detection and modeling

This dissertation proposes methodologies to detect and model activity episodes from WiFi traces. Its contributions can be categorized as *activity-episode sequence detection*, *activity*

path choice modeling and *location choice modeling*. The different steps to reach a complete activity-based pedestrian demand model are presented in Fig. 1.1.

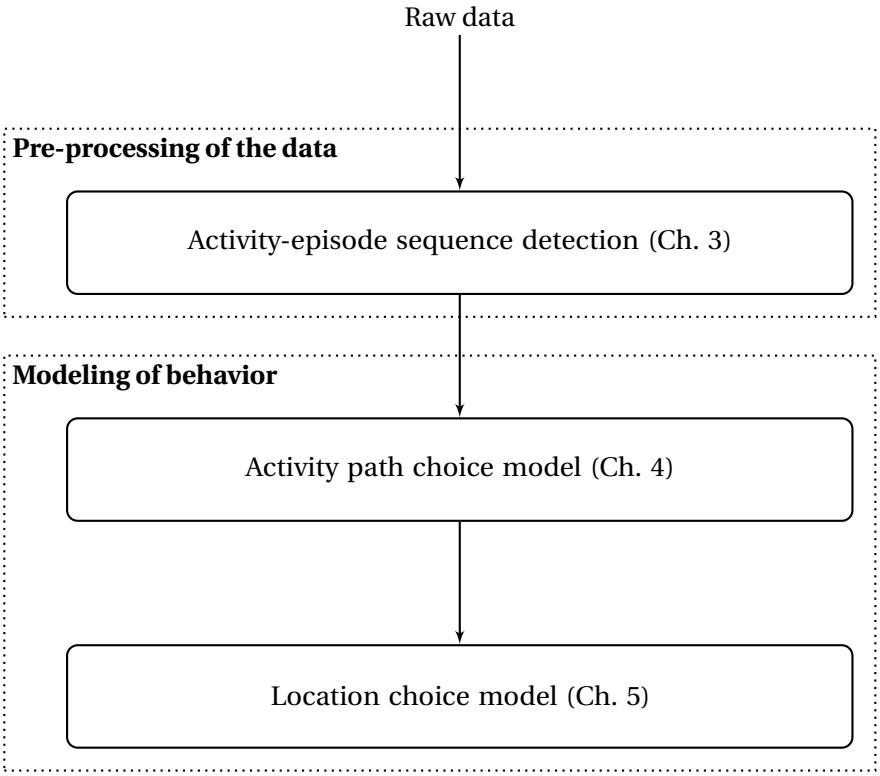


Figure 1.1 – A full system for the modeling of strategic behavior from raw localization data, including pre-processing of the data and modeling of activity and destination choice models.

1.3.1 Activity-episode sequence detection

WiFi traces offer the opportunity to collect panel data in the long term, but do not directly detect activity-episode sequences. In Ch. 3, stop and activity at the stop are extracted by merging the WiFi localization data with land use information.

Explicitly modeling the imprecision in the measure Our Bayesian approach takes into account the fact that pedestrian networks are usually denser than other mobility networks and localization is often sparse, in particular indoors. This methodology is robust for low density measurements. Ambiguity is explicitly stated through the likelihood of each activity-episode sequence.

Using prior knowledge of the infrastructure The network traces of a device are supported by the *a priori* knowledge of the underlying pedestrian map and attractivity of the activities. Time constraints—such as schedules for trains in a railway station, for planes

in an airport, for concerts in a music festival, or for classes on a campus, or opening hours for shops or restaurants—can be added to the model. The usage of a pedestrian network corrects for anisotropy in the facility.

Avoiding the pingpong effect If access points are changing very often from one to another while the device is in fact static, the true activity location does not change with our approach.

We present an empirical study on a campus.

1.3.2 Activity path choice model

In Ch. 4, we describe a model for the choice of an activity pattern. We focus on choice set generation of activity patterns using recent developments in route choice modeling and strategic sampling.

No tours, no priorities Our modeling approach is not structured on home and tours from home, since it is not adapted to the public space context. It does not assume any priorities between activities or episodes, primary and secondary activities, or mandatory and discretionary.

Managing large choice sets Modeling simultaneously the choice of activity type, time of day and duration generates large choice sets. The large dimensionality of the choice set is managed through strategic sampling using a Metropolis-Hastings algorithm.

Unique utility for activity type, time-of-day and duration Focusing on simultaneous choices of activity type, duration and time-of-day, the chosen alternative is one sequence of activity episodes. Represented as a path in a network, the sequence is a single choice and utility is associated with the full pattern.

1.3.3 Location choice model

Chapter 5 describes a location choice model conditional on the activity-episode sequence.

Including panel data Chapter 5 specifically develops a modeling framework to account for panel data in location choices. It allows to understand people's habits in their decision processes

Correcting for serial correlation Only a few dynamic models of location choice exist in the literature, and none of them to our knowledge correct for serial correlation. We apply the Wooldridge (2005) method to deal with the initial values problem on the choice of catering location on EPFL campus using WiFi traces.

We present cross-validation results. We also present elasticities to price and forecast the scenario of the opening of a new catering location. Predicted market shares of the new catering location correspond to point-of-sale data of the first week of opening.

1.4 Outline

Chapter 2 reviews the literature in pedestrian data, activity-episode detection, activity and location choice modeling.

This thesis is then structured in two main parts:

Part I presents the preprocessing of WiFi traces in order to detect stops and activity type at the stops. It provides activity-episode sequences from raw localization data.

Chapter 3 proposes a Bayesian approach using WiFi traces to detect pedestrian activity-episode sequences. It merges WiFi traces with attractivity measures, map information and time constraints. A case study on the EPFL campus is presented, with validation and sensitivity analysis.

This chapter has been published as:

> Danalet, A., Farooq, B., and Bierlaire, M. (2014). A Bayesian Approach to Detect Pedestrian Destination-Sequences from WiFi Signatures, *Transportation Research Part C: Emerging Technologies* 44: 146-170.

Part II focuses on modeling the activity and location choices from the activity-episode sequences.

Chapter 4 first describes the general framework for modeling activity-episode sequences, decomposing the behavior into activity and location choices. The probability of reproducing the observation of a sequence of measurements is formally expressed.

Then, this chapter proposes a model for the simultaneous choice of activity types, order, start times and durations of activity episodes in a sequence. In particular, we develop a framework for choice set generation based on path choice to deal with the large choice set. A case study using WiFi traces on EPFL campus is presented.

This chapter has been submitted for publication in a special issues of an academic journal for the 14th International Conference on Travel Behaviour Research.

> Danalet, A., Bierlaire, M. (2015). Strategic sampling for activity path choice.

Chapter 5 applies the Wooldridge (2005) method to deal with the initial values problem on the choice of catering location on EPFL campus using WiFi traces. Cross-validation and forecasting in the scenarios of cost increase and opening of a new catering location are presented.

This chapter has been submitted for publication in the Journal of Choice Modeling:

Danalet, A., Tinguely, L., de Lapparent, M., Bierlaire, M. (2015). Location choice with longitudinal WiFi data. Technical report TRANSP-OR 151110.

Chapter 6 concludes the thesis, review theoretical and policy applications, together with limitations of the present study and future research directions.

2 Literature review

In this literature review, pedestrian data are first described (Section 2.1). Then, we present the different strategies used in the literature to transform raw data into activity episodes (Section 2.2), to model the activity choice (Section 2.3) and the destination choice (Section 2.4).

2.1 Pedestrian data

The recent developments in detection technologies open doors to new research about pedestrian behavior. Traditional data are collected by manual counting, mechanical counting, infrared beans and video surveillance. We briefly introduce the purpose, the scale and the context of these data collection campaigns, and review the opportunities and drawbacks of these traditional tools and of recent technologies. Then, we further develop on mobile phone tracking data (from an antenna perspective) and map data in Section 2.1.1 and 2.1.2.

Purpose Data about pedestrians are collected for health (e.g., Shephard; 2008; Ding and Gebel; 2012; Sugiyama et al.; 2012), security (e.g., Candamo et al.; 2010; Popoola and Wang; 2012), marketing and sales (e.g., Kholod et al.; 2010; Hui et al.; 2013; Yaeli et al.; 2014) and transportation purposes. In transportation, collected data are used for safety, security, efficiency and attractiveness (Bauer et al.; 2009), for modeling activity, destination, mode and route choices and walking behavior (Bierlaire and Robin; 2009), at strategic, tactical and operational levels (Hoogendoorn and Bovy; 2004).

In practice in transportation, these data are used to plan evacuation in case of emergency or model normal behavior, both in cities or in pedestrian facilities, to provide travel guidance and information helping the pedestrians in implementing their journeys, and to build efficient facilities, and manage them on a daily basis.

In this thesis, we focus on transportation purposes, in particular the strategic level of modeling activity and destination choices in a normal, non-evacuation context. The goal of such models is to understand the demand for different points of interest in the pedestrian facility and

forecast demand management policies, e.g., about the localization of points of interest or the impact of schedules on the demand for the different categories of points of interest. There is no specific literature about the data requirements for such models in the pedestrian context.

Scale Pedestrian data collection campaigns take place at city or pedestrian facility scales. We motivate this decomposition in two levels by the presence or absence of mode choice. Recent attempts centralize pedestrian traffic counts at the national level (Nordback et al.; 2015), but as a decision aid tool for cities.

At the city scale, information on pedestrian traffic is an "emerging area" as mentioned by the Traffic Monitoring Guide (TMG) of the US Federal Highway Administration (FHWA; 2013). It includes information on monitoring pedestrians for the first time in 2013. Apart from traffic flows, data about walkability are also collected at the city scale (Pivo and Fisher; 2011) and household travel surveys begin to include walk as a travel mode (Berge and Peddie; 2010; Morency et al.; 2011; Millward et al.; 2013; Ravalet et al.; 2014).

Pedestrian facilities include train stations, airports, concert halls or festivals, shopping malls, campuses, hospitals, sports infrastructure such as stadiums, or even religious infrastructures. In these facilities, people are supposed to walk and other transport modes are exceptional. Originally, pedestrian data are collected at this scale mostly for safety issues, e.g., in Mecca (Helbing et al.; 2007) or for music festivals, like for the Love Parade in Duisburg (Helbing and Mukerji; 2012; Krausz and Bauckhage; 2012). In a normal behavior context, studies have been performed in train stations (Daamen; 2004; Ueno et al.; 2009; Hänseler et al.; 2014; Ton; 2014; van den Heuvel et al.; 2015), airports (Manataki and Zografos; 2009; Solak et al.; 2009; Wu and Mengersen; 2013; Kalakou et al.; 2014; Liu, Usher and Strawderman; 2014), hospitals (Yao et al.; 2011; Khan; 2012; Prentow et al.; 2014), music festival (Naini et al.; 2011; Versichele et al.; 2012; Duives et al.; 2014), museums (Kanda et al.; 2007; Lanir et al.; 2014; Yoshimura et al.; 2014) and commercial centers (Zhang et al.; 2012; Yaeli et al.; 2014).

Depending on the purpose and the scale, different data collection techniques are used. When data are related to the full facility, collection techniques are revealed and stated preferences surveys (Kalakou et al.; 2014; Liu, Usher and Strawderman; 2014), based on the shape of the built environment (Ueno et al.; 2009; Khan; 2012; Zhang et al.; 2012) and/or aggregate demand such as train or flight schedules (Manataki and Zografos; 2009; Solak et al.; 2009), or using wireless technologies such as WiFi, Bluetooth or RFID (Naini et al.; 2011; Yao et al.; 2011; Versichele et al.; 2012; Prentow et al.; 2014; Ton; 2014; Yaeli et al.; 2014; Yoshimura et al.; 2014); when focusing on a specific area such as a corridor and on more microscopic behaviors, data must be more precise and different technologies are used, e.g., visual, depth and infrared sensors (e.g., Hänseler et al.; 2014) or laboratory experiments (e.g., Moussaïd et al.; 2009). In this thesis, we focus on the scale of pedestrian facilities.

Data collection techniques As seen in the previous paragraph, when trying to understand people's behavior in a pedestrian facility, one can directly ask the person (revealed and stated preferences surveys), look at the map, study flight or train schedules, use wireless technologies or directly observe the person (visual, depth or infrared sensors). We review different techniques here.

Data are traditionally decomposed in two main types: counts and trajectories (Bauer et al.; 2009; Bierlaire and Robin; 2009). Counts are further decomposed into intersection or segment counts (Nordback et al.; 2015), while tracking methods are decomposed between intrusive and non-intrusive in Bauer et al. (2009), i.e., when distribution of a tracking device is needed or not. We briefly describe the different data collection methods for counting and tracking below, and discuss their advantages and disadvantages.

Manual counts and shadowing Manual counts are performed by individuals, in the field or on a video recording. Data usually cover a short duration at infrequent intervals and are biased by weather conditions, events, and weekly and seasonal variations (Nordback et al.; 2015). It is the most expensive way for counting pedestrians and a common practice (FHWA; 2013; van den Heuvel and Hoogenraad; 2014). Human observers usually underestimate the flows by 8 % to 25 % (Diogenes et al.; 2007). Depending on the complexity of the scene and the motivation of the human observers, counting on a video recording is not necessarily more accurate than field work, but can be repeated (Greene-Roesel et al.; 2008).

The tracking counterpart of manual counting is shadowing (or stalking). It has been used in many studies (e.g., Routledge et al.; 1974) and in particular performed in train stations (Daamen; 2004; Millonig and Maierbrugger; 2010). It is mostly used for qualitative research, in sociology or ethnography (Quinlan; 2008), since it is difficult to collect a large number of traces.

Mechanical counts Mechanical counts are automated and usually cover a longer duration than manual counts (Bauer et al.; 2009; Nordback et al.; 2015). They are usually considered as cheaper than other counting sensors (Bauer et al.; 2009). They include pressure and seismic sensors (pressure mats), sensitive to the weight of the pedestrians, and turnstiles.

Survey data Revealed preference (RP) data about mode or location choices are collected using household travel surveys (Atasoy et al.; 2013; Ravalet et al.; 2014) or on street/in pedestrian facilities surveys (Kalakou et al.; 2014). Household travel surveys usually focus on long and motorized trips and regroup cycling and walking in one "soft mode" category (e.g., Atasoy et al.; 2013). Some examples do include walking and short trips, in order to study number of steps (Morency et al.; 2011) or typology of walkers (Ravalet et al.; 2014). Recall questionnaires ask the respondent to remember a past decision and report about it. Liu, Usher and Strawderman (2014) typically ask the respondents to describe their last visit in an airport. They are not

very precise and face recall bias and social-desirability bias. Millward et al. (2013) suggest to use GPS-assisted prompted recall survey to determine frequency, length and purpose at destination of walking episodes. One issue with revealed preference surveys is the difficulty to collect a lot of observations (Liu, Usher and Strawderman; 2014 use 400 activity patterns with a web-based revealed preference survey in an airport).

Lasers and infrared sensors Infrared Sensor (Active or Passive) is a common practice and is less expensive than manual counting, according to FHWA (2013). It can also be used for tracking pedestrians, but cost per area is high (each sensor only monitors a few square meters) (Bauer et al.; 2009). Challenges related to such technique include detecting the walking direction and dense scenarios when occlusion happens (Bauer et al.; 2009), in particular high errors with groups (FHWA; 2013). Laser range scanners have the same problem with dense crowds. They can be used for counting and tracking, by sensing the distance to the nearest object. Laser scanners are expensive compared to video cameras (Bauer et al.; 2009).

Video counting Video counting is the usual data collection technique used in a crowded environment (Versichele et al.; 2012). The occlusion issue is usually managed by using a top view (Bauer et al.; 2009). Lighting can cause problems, as well as weather (Bauer et al.; 2009; FHWA; 2013). Specifically for counting, the technology has potential accuracy in dense, high-traffic areas, but counting algorithm development is still maturing, video counting underestimate count flows in dense conditions (Bauer et al.; 2009) and "much of this university research has not been incorporated into existing commercially available products" (FHWA; 2013). For trajectory detections, Alahi (2011) and Alahi et al. (2014) use networks of cameras to track and analyze pedestrian trajectories. Alahi's main motivation is the number of already installed cameras generating large datasets. Finally, video image processing typically has the highest equipment costs (Bauer et al.; 2009; FHWA; 2013).

GPS, RFID tags and readers, smart card, ... Passive RFID tags, reacting to the signals of a RFID reader, can be used for counting. They are mostly used for access controls in offices, museums, etc. (Bauer et al.; 2009). Active tags with their own power supply could possibly be used for tracking (Bauer et al.; 2009). Hendrich et al. (2008) use RFID to identify how nurses spend their time on specific activities and the distance they travel in 36 hospitals. Kholod et al. (2010) use RFID to collect the shopping path length in grocery store.

GPS devices are precise but raise the issues of distributing and recollecting the devices and of the precision indoor and in dense urban settings (Versichele et al.; 2012).

van den Heuvel and Hoogenraad (2014) use *smart card data* (or automatic fare collection (AFC) data), i.e., check-in and check-out of passengers, in a train station. These data provide information on the speed and the route of passengers in the train station. RFID tags used for ticketing can also be used without an explicit check-in or check-out by the user. Yu Quan et al.

(2011) use such sensors to detect the speed of pedestrians. Such data are very recent and not yet used in practice for tracking pedestrians at a large scale.

Data from the mobile phone Location based-services provided directly by smartphones generate relevant data. A large proportion of the population carry and use their smartphone in everyday life and sensors into mobile phones offer large quantities of data for ubiquitous observations of their owner. The development of research based on these data is due to

1. the number and quality of sensors in recent phones,
2. the possibility to create third-party apps and easily install and deliver them on the user's phone and
3. the easy transmission of collected data allowed by mobile data (Lane et al.; 2010).

The device's owner can explicitly provide data input (*participatory sensing*) or data can be collected autonomously without user involvement (*opportunistic sensing*) (Hoseini-Tabatabaei et al.; 2013). Participatory sensing faces the issue of bias related to user's opinion (McNeill and Chapman; 2005; Hoseini-Tabatabaei et al.; 2013). Mobile phones contain inertial, ambient and positioning sensors. We review here only positioning sensors (i.e., GSM, GPS, WiFi, Bluetooth). A review of inertial and ambient sensors can be found in Hoseini-Tabatabaei et al. (2013).

Cell tower signals are used by the phone to call and transfer data. Collecting the GSM cell ID (CID) identifies the base transceiver station (BTS) and localize the mobile user. The fluctuation pattern of cell IDs associated with signal strength provide information on user's mobility (see Anderson et al.; 2007 for an example in the health context).

GPS is more precise than cell tower signals. It can be used for pedestrian navigation (Arikawa et al.; 2007). Nevertheless, it works only outdoor. It also reduces the battery life of the phone (Gaonkar et al.; 2008). The battery life issue is usually dealt with by using other sensors, such as WiFi or cell tower signals.

A mobile phone can locate itself when connecting to WiFi by knowing the location of access points. Due to the large signal transmission range, the positioning accuracy is low, and signal strength, signal triangulation and fingerprinting can be used to improve the localization (when there are more than one access point available). WiFi is the second most power-demanding sensor after GPS (Gaonkar et al.; 2008) for providing location information. Similarly to cell tower, it can be used indoor.

Bluetooth is designed for short-range communication. A phone can detect other devices using Bluetooth, uniquely identify them through Bluetooth MAC address (also called Bluetooth identifier - BTID), detect the device type and the device name (Hoseini-Tabatabaei et al.; 2013).

All data from the mobile phone require the explicit agreement of the device's owner and generate privacy issues. They also require the transmission of the data to the analyst, either using WiFi when available or GSM or 3G networks otherwise. In (Ball et al.; 2014), 59 % of the participants of a smartphone-based travel survey use GSM or 3G network for the upload of their data. 3G data was of greatest battery cost than GPS and WiFi location information.

Ultra-sonic measurements, radar measurements, pedometers are not reviewed here, as they are not often used in the transportation literature. Pedometers have been mostly used for health related issues (Bierlaire and Robin; 2009). A review of the existing data collection techniques show that data mostly focus on local behavior (Papadimitriou et al.; 2009). Data collection techniques face issues such as small sample sizes (manual counts and shadowing, surveying), cost (manual counts and shadowing, video counting), underestimation of counts (manual counts), precision (surveying), recall and social-desirability biases (surveying), the need to distribute devices (e.g., GPS devices, RFID tags) and the need for agreement of the user and the privacy risk of accessing other data with data from the mobile phone. In the case of smart card data, only passengers are tracked and not all visitors in the train station.

In the following sections, we focus on mobile phone tracking data from antennas, often existing in pedestrian infrastructure and covering the full infrastructure (Section 2.1.1), and on map data (Section 2.1.2).

2.1.1 Mobile phone tracking data

Similarly to Alahi's motivation about pre-existing network of cameras (Alahi; 2011; Alahi et al.; 2014), networks for mobile phones already exist, a majority of people are carrying a mobile device such as a smartphone, and they generate data. Data from communication network infrastructure ("network traces") include all data from antennas providing communication to the mobile phone, i.e. cell tower data, WiFi access point data and Bluetooth sensors. It does not include GPS, since there is no access to the satellites data. According to Calabrese et al. (2014), mobile phone tracking data are used for estimating population distribution, activity types and mobility patterns, analyzing local events and the geography of social networks.

Using traces from communication network infrastructure has several advantages on data from the smartphone. First, full coverage of the facility/area of interest is usually cheap (Versichele et al.; 2012) and allows for an estimation of the overall demand (Yoshimura et al.; 2014). The communication infrastructure sometimes already exists, and increasing its density has a positive side effect. Smartphone users do not need to install anything on their device, and so the access to sensitive information such as emails or address book is limited for the analyst, which ensure privacy for the users. There is no need to distribute and recollect tracking units, network traces usually generate large samples of data, there is no bias related to the individual being aware of being followed, and the different technologies (Bluetooth, GSM data, WiFi) work indoor (Versichele et al.; 2012; Yoshimura et al.; 2014). Finally, traces from communication network infrastructure are related to the infrastructure and not to the

individual: all individual smartphones going through a facility are tracked and not all places visited by the same individuals. It allows the analyst to focus on the pedestrian facility/area of interest covered by the communication network.

There are few drawbacks to network traces as well. Socio-economic and demographic attributes are difficult to collect due to both privacy concerns (if the data already exist) and to the difficulty to survey the tracked person from the infrastructure side (if the data does not exist). Additionally, smartphone users are not necessarily representative of the full population. Smartphones' users might represent a biased sample of the population, in particular under-representing elderly people (Versichele et al.; 2012; Calabrese et al.; 2013). Network traces have the advantage of long data collection duration. Nevertheless, compared to data from the mobile phone, the frequency of measurements in network traces depends on the usage of the phone and is often related to an action by the user (making a call, using internet, receiving an SMS, etc.).

Indoor localization is complicated due to walls and different obstacles, blocking waves propagation and reflecting signals. Different techniques are developed to mitigate the measurements errors, including triangulation (lateration techniques, methods based on signal attenuation and Received Signal Strength (RSS), etc.), scene analysis such as fingerprinting, or proximity measures. Without entering the technical details, let's mention that fingerprinting, i.e., constructing a radio-map by collecting signal strength samples, has been criticized for being time consuming, labor intensive, vulnerable to environment changes and expensive (Hossain and Soh; 2015). Fingerprinting for a large area is impractical and calibration-free methods are preferred ("large" being more than 600 m^2, since Faragher and Harle (2015) performed fingerprinting for such a surface). A review of positioning algorithms and technologies can be found in Liu et al. (2007). We briefly summarize some observations on different technologies below.

Technologies

Mobile phone tracking data include *mobile phone network data* (also called *GSM data* or *cellular telephone signals*). Mobile phone network data from cell towers are the equivalent of cell tower signals from mobile phones, from an antenna point of view. They are provided by the phone operators (in Estonia, Ahas et al.; 2010, in the US, Isaacman et al.; 2011, in Côte d'Ivoire (Ivory Coast), Liu, Janssens, Cui, Wang, Wets and Cools; 2014; Palchykov et al.; 2014). Call data records (CDR) are only generated when the device is in use. The only option to get continuous cellular tower data consists in installing a logging app on the mobile device (Eagle et al. 2009). These data have been used to detect home and work locations (Ahas et al.; 2010; Isaacman et al.; 2011). Mobile phones can be tracked in most indoor contexts, but only with a precision of several hundred meters in practice (Yaeli et al.; 2014) (50-200 m depending on cell size according to Liu et al.; 2007, but it can be much more, in particular in low density areas). A complete review is available in Calabrese et al. (2014).

The range of *Bluetooth scanners* vary between 10 to 100 m (Versichele et al.; 2012), typically 10-15 m (Liu et al.; 2007). The penetration rate varies between 7 % and 11 % (Versichele et al.; 2012; van den Heuvel et al.; 2015). Bluetooth traces have been used to study spatiotemporal crowd density variations, typology of visitors, the visit duration, and the flows (Versichele et al.; 2012; Yoshimura et al.; 2014), the transport mode to reach the zone (Versichele et al.; 2012) and route choice (Yoshimura et al.; 2014; van den Heuvel et al.; 2015).

Wireless local area networks (WLAN) allows to track devices. The range of WLAN is 50-100 m (Liu et al.; 2007). WiFi traces are collected indoor, e.g., in stores (Yaeli et al.; 2014) or for conferences (Krueger et al.; 2015). The device (e.g., a smartphone) does not technically need to be logged on to the WiFi network, but just to have WIFi antenna turned on. Identification is made through the MAC address and some vendors started recently to randomize it, but the real MAC address is used when the device connects with the WiFi network (Yaeli et al.; 2014). The typical accuracy of WiFi positioning systems using Received Signal Strength (RSS) is 3 to 30 m (Liu et al.; 2007).

Mobile phone network data are too imprecise to be used at the scale of a pedestrian facility (Yaeli et al.; 2014). Compared to GSM data, Bluetooth tracking is more precise (Yoshimura et al.; 2014) but still has a low penetration rate (Yaeli et al.; 2014). Near field communication (NFC) is included in many phones but has a maximum range of about 5 cm. Both Bluetooth tracking and NFC require to install scanners, which is not needed with WiFi access points. In this thesis, we propose to detect pedestrians from already deployed WiFi network infrastructure, originally not designed to provide such data, using commercially available WiFi tracking tools without performing fingerprinting (see Section 3.4.1). Even when installing sensors, this approach is cost-effective for large pedestrian facilities "with continuous operational challenges or redevelopment activities" (van den Heuvel and Hoogenraad; 2014).

2.1.2 Map data

This section focuses on the information in pedestrian networks. Existing pedestrian networks are described. Then, space syntax, a possible usage of information from pedestrian networks, is briefly described and criticized.

The pedestrian network depends on the scale of the study area and on the definition of destinations: buildings (Yoon et al.; 2006), WiFi access points (APs) (Wanalertlak et al.; 2011; Tuduce and Gross; 2005), rooms (Sevtsuk et al.; 2009), or, at an urban scale, subzones of the motorized regional zone system as nonmotorized destinations (Eash; 1999).

The walking distance between the destinations is usually not available (Kasemsuppakorn and Karimi; 2013). Yoon et al. (2006) converted a map to a graph between buildings and limited themselves to major roads. In the extension of a Chicago model for nonmotorized trips, the Manhattan distance is motivated by the grid plan and the absence of a pedestrian network (Eash; 1999).

Kasemsuppakorn and Karimi (2013) propose to build the pedestrian network from GPS traces, which does not work indoor. Kang et al. (2004) are using WiFi to cluster places of interest and label them, but APs serve different kinds of locations surrounding them (Calabrese et al.; 2010). Without a model based on a pedestrian network, changes in the pedestrian facilities such as pedestrian bridges or underpasses cannot be tested.

Indoor networks of pedestrian facilities allowing for computation of the shortest path between two destinations are increasingly available for airports, museums, campuses, hospitals and malls due to the complexity of path finding (Goetz and Zipf; 2011). Crowdsourced geodata such as OpenStreetMap are extending to indoor spaces (Goetz; 2012).

On campuses, University of Ottawa proposes a shortest path using indoor routes depending on shortest or "warmest" options[1]. Of particular relevance to the application within this thesis, EPFL website proposes an orientation tool for the campus, http://map.epfl.ch. It allows localization of offices or any place in the university, as well as itineraries between these places. Moreover, it offers thematic maps, such as lists of restaurants. Created in 2002, it is based on a student project (Büchel; 2004). From Autocad files (i.e., drawings), new data were generated: a network, with vertices and edges in 3D, with lifts, stairs, ramps and doors. This process used Feature Manipulation Engine (FME) scripts. This tool results from a collaboration between Camptocamp, an EPFL spin-off company, EPFL's Real Estate and Infrastructures Department (DII), EPFL's Information Technology Domain (DIT) and EPFL's Knowledge and Information Services (KIS). It is based on PostgreSQL and PostGIS as databases and Mapserver as visualization tool (Philipona; 2002).

One possible usage of pedestrian network in understanding pedestrian behavior is space syntax (Hillier and Hanson; 1984; Hillier; 1999). Space syntax quantifies the 2D spatial configuration of the built environment. As stated by Ratti (2004), space syntax is "an extension of network analysis concepts into architecture and urban planning". Space syntax has been used to evaluate indoor building configurations, in hospitals (Haq and Luo; 2012; Khan; 2012), airports (Kalakou et al.; 2014) and museums (Choi; 1999; Hillier and Tzortzi; 2006). The spatial and visual layout of the rooms and corridors are described with measures of accessibility, depths (i.e., number of changes in direction) of the pedestrian network, measures of connectivity of the different rooms, visual accessibility (i.e., the number of points visible from a given location). Space syntax has been criticized as not being able to model pedestrian choice making (Ratti; 2004). It does not consider street size and length (Ratti; 2004; Haq and Luo; 2012), building height, land use (Ratti; 2004), or task or motivational requirements (Lu et al.; 2009; Haq and Luo; 2012). As we will see in this thesis, distance is a main driver for destination choices of pedestrians (see Section 5.3.1), which contradict the discarding of metric information by space syntax. In this thesis, we propose to use pedestrian map information, including metric information, and merge it with network traces and land use data (see Ch. 3).

[1] http://www.uottawa.ca/maps/

2.2 Activity-episode detection

2.2.1 From diary surveys to location-aware technologies

One recent trend in travel demand modeling is the usage of data from location-aware technologies (Chen and Yang; 2014; Danalet et al.; 2014; Miller; 2014; Carrel et al.; 2015). Traditionally, collected data are revealed preferences about activity and travel patterns from diary surveys, where people describe 1-2 past days (Ettema; 1996; Carrel et al.; 2015). The largest panel surveys include a six-week period for 317 participants (Axhausen et al.; 2002), a six-week period for 261 participants (Axhausen et al.; 2007) and a twelve-week period for 71 participants (Schlich; 2004). Most long-term surveys cover a maximum of 7 days and are not panel data (Ortúzar et al.; 2011; Carrel et al.; 2015). GPS-based prompted recall activity-travel survey allows for longitudinal surveys, using GPS devices carried by respondents (Frignani et al.; 2010; Yang and Timmermans; 2015). Recall methods can be implemented on mobile devices (Rindfüser et al.; 2003; Cottrill et al.; 2013).

Location-aware technologies help improving the quality of explicit surveying. They can also be used alone, from the communication infrastructure side, such as cell tower traces or WiFi access points traces (Bekhor et al.; 2013; Calabrese et al.; 2013), or from the individuals' devices (Etter et al.; 2012; Buisson; 2014; Chen and Yang; 2014; Carrel et al.; 2015). Etter et al. (2012) show that it is possible to predict up to 60% of next visited places from passive smartphone data.

2.2.2 Preprocessing: stop and semantics detection

Data preprocessing methods are needed to transform these raw observations into data adapted for modeling purpose. First, detection of stops points discriminates places where people spend time and perform activities from moving between these stop points (Rieser-Schüssler; 2012; Jiang et al.; 2013). After cleaning the data, Bekhor et al. (2013) define a destination as a cell tower where the device is connected for more than 20 minutes without changing. Using triangulation from cell towers, Calabrese et al. (2013) merge all measurements in a time interval ΔT with maximum distance 1 km. ΔT is not given. In Jiang et al. (2013), the first step is similar, with a distance of 300 m and a stay time of 10 min. Here, the accuracy of the location is about 200 to 300 m. The second step associates with each other the different stop points at different times if they are close using a grid-based clustering method. It allows to identify the places that are visited multiple times, despite measurement errors. Based on triangulation from WiFi access points, Danalet et al. (2014) associate all measurements with points of interest (POI) in the map, compute the travel time between these POI, define a distribution for arrival and departure time based on travel time and measurement timestamp, and finally remove destinations with expected duration smaller than $T_{min} = 5$ min. From WiFi data from smartphones, Buisson (2014) clusters measurements using a Density-based Spatial Clustering of Applications with Noise (DBSCAN) algorithm.

Pure location-aware technologies lack the path semantics (Miller; 2014). To overcome this issue and detect activity purpose, localization data are merged with land-use information (Rieser-Schüssler; 2012; Miller; 2014; Danalet et al.; 2014). A review of research in this direction before 2012 is available in Miller (2014). Jiang et al. (2013) propose a visual example of how land use data could be applied, without a general methodology. In a pedestrian facility, we propose a Bayesian approach in Chapter 3. In urban context, Buisson (2014) also uses a Bayesian approach. For a given cluster of access points:

$$P(\mathscr{A}_k|\hat{t}) \propto P(\hat{t}|\mathscr{A}_k) \cdot P(\mathscr{A}_k) \tag{2.1}$$

where $\mathscr{A}_k$ is an activity type and $\hat{t}_{1:J}$ is a set of measurement timestamps corresponding to the cluster. $P(\hat{t}|\mathscr{A}_k)$, the probability of generating a signal at a certain time knowing the activity type, is computed using time-use statistics, e.g., from travel diary surveys. The prior is similar to the one defined in Ch. 3, using OpenStreetMap data for list of POI, and census and national statistics for number of residents and employees.

Several applications using data from communication infrastructure have been developed to study mobility behavior, both with WiFi (Yaeli et al.; 2014) and GSM traces (Bekhor et al.; 2013; Calabrese et al.; 2011). These new data collections are motivated by the needs for calibrated agent-based models. Post-processing methods are needed to transform these raw observations into data adapted for modeling purpose to overcome imprecision and missing observations in the data: detection of stops points, activity purpose detection through land-use information and spatial matching (Rieser-Schüssler; 2012). Hoseini-Tabatabaei et al. (2013) review these needs for mobile phone sensing. With GSM traces, Bekhor et al. (2013) mention the elimination of "unreasonable movements performed in short time periods between antennas located far apart" without more details. Calabrese et al. (2011) does not consider the underlying transportation network to correct for anisotropy.

2.2.3 Mobility models for WiFi traces

A large literature exists about WiFi traces from a computer communication point of view. A complete review can be found in Aschenbruck et al. (2011). All references in this paper define mobility trace-based models as a tool to improve the quality of the WiFi. The goal is to predict the next point-of-attachment of the user (Wanalertlak et al.; 2011). This body of work studies mostly pedestrians, since the scale of the problem is related to offices or campuses. Yoon et al. (2006) study mobility models for WiFi infrastructure and try to make them representative of real mobility, and mention the possible applications to urban planning, socially-based games, or augmented reality. Field studies have been done (Tang and Baker; 2000; Balachandran et al.; 2002; Balazinska and Castro; 2003; Yoon et al.; 2006; Sevtsuk et al.; 2009; Zola and Barcelo-Arroyo; 2011; Wanalertlak et al.; 2011; Meneses and Moreira; 2012). The main results are the prediction of changes in access points (APs). The main problem reported in these articles is the *ping pong* effect, when the device has similar signal strengths from different APs and changes regularly from one to another. This is a problem from a network viewpoint, and

also for modeling pedestrian origins and destinations. Yoon et al. (2006) propose to use a moving average weighted by time spent at destination to remove the extra AP logs. A general solution presented in Aschenbruck et al. (2011) consists in aggregation of data over different APs. Most studies about WiFi are focusing on network performance and management and not on human mobility. In Yoon et al. (2006), contrarily to all other papers cited here, an OD matrix is estimated at the building level in Dartmouth college. Variations in time/day are not considered, as Aschenbruck et al. (2011) noticed.

In the literature about mobility models for WiFi infrastructure from a computer communication point of view, the most common model, Random Waypoint model (RWP), is often criticized as not representing real human mobility (Conti and Giordano; 2007). One of the problems with RWP consists in using straight lines between two signals in different APs, even if this path is not physically possible. This is the main reason why trace-based mobility models were developed in this domain of research. A key challenge in building a realistic model is to define a pedestrian network and the corresponding possible paths that the user with a device can follow. This process of constructing and using the pedestrian network in order to improve the mobility model is not explicitly presented in Aschenbruck et al. (2011) in their large review of trace-based mobility models. The need for a more complex approach is emphasized in Rojas et al. (2005).

2.3 Activity choice modeling

2.3.1 Early works

Before 1950, transportation studies didn't have predictive power and mostly described the current state of traffic (Weiner; 1999). In 1954, Mitchell and Rapkin (1954) theorized models for travel patterns and behavior based on trips and susceptible to change depending on attributes such as land use or socioeconomic characteristics. They already mentioned at the time the limitations of a trip-based approach to really understand motivations of travel.

In practice, travel demand analysis has been decomposed in four sequential steps in the Urban Transportation Modeling System (UTMS): trip generation, trip distribution, modal split and traffic assignment. The fundamental unit of these analysis is a trip. Trips are generated not from a behavior-based demand but from a physical analogy with gravity: from trip production location to trip attraction location. In 1955, the Chicago Area Transportation Study (CATS) used this analytical decomposition.

In the 1970s, a shift was observed from the traditional trip-based approach to a demand-oriented approach. Trips started to be considered as a derived demand. Scheduling decisions were surveyed, e.g., in the Household Activity Travel Simulator (HATS) (Heggie and Jones; 1978; Jones; 1979). At the same time, models started to be more of a disaggregate microsimulation approach. This change in paradigm is the origin of activity-based travel demand modeling. It is related to the rapid development of random utility-based individual choice models for route,

mode, frequency and destination choice (McFadden; 1974; Stopher and Meyburg; 1975).

The premise of activity-based approach consists in considering activity as a choice and trips as a way to complete the chosen activity. In other words, modeling the daily activity patterns allows the development of behavioral travel demand models that are sensitive to changes in policy.

For Hägerstraand (1970) the choice set of activities is constrained in three ways:

Capability constraints are biological or technological constraints of the individual. Some are fundamentals and structure mostly the time when activities are performed, mainly sleeping at night and eating at some time of the day. Others are distance oriented and express the time-space constraints on movement. People need to perform a trip to reach a destination before being able to perform an activity.

Coupling constraints define the requirement for some activities of other people or resources (salesman and customer in a shop, students and teacher in a class, meetings at work).

Authority constraints are protecting resources and limit accessibility to certain persons. They can be related to payment, invitation, power or custom: a seat in a theater is a temporary authority constraint, while traffic rules are permanent.

In practice, there are traditionally two main components in activity-based models: activity generation (dealing with the basic needs for self-realizations) and activity scheduling (spatio-temporal constraints and opportunities related to the actual activities). The interactions between these two components are in both directions.

Activity generation is facing the problem of generating the choice set for activity patterns and different techniques have been proposed. Using discrete choice models, Adler and Ben-Akiva (1979) proposed one of the earliest examples of these models. The patterns are chosen on attributes such as modes, number of destinations, purposes, time spent at destination, travel times, etc. Bowman and Ben-akiva (1996) propose a hierarchical approach, with a decomposition into primary and secondary tours, assuming some activities are more important in structuring the travel patterns. For both types of tours, they model the activity pattern, the choice of tour and the choice of destination itself.

Another stream of research assumes that people are not rational utility maximizers and do not always take optimal decision. This type of *rule-based* models are numerous (Hayes-Roth and Hayes-Roth; 1979; Jones et al.; 1983; Pendyala et al.; 1998; Arentze and Timmermans; 2004, among others). Based on the works by Newell and Simon (1972) and Tversky and Kahneman (1981) in psychology, sub-optimal decisions are related to a satisficing approach or a limitation in information acquisition and treatment. Habib (2007) reviews these models and mentions some of their limitations. They are good for modeling short-term policy analyses but not for long-term demand forecasting and they need specific rules to be able to response to changes

in policy. Inputs are generated empirically and lack theoretical foundation. Handling of preference heterogeneity is a difficult task in these models. Bowman and Ben-akiva (1996) criticized these models and the absence of dependencies in activity choices across the day.

2.3.2 Contemporary approaches

This review presents some relevant aspects of activity choice for our approach. Concepts from route choice literature are also used in this thesis and are briefly reviewed here. General reviews about activity-based travel demand modeling can be found in Ettema (1996), Bhat and Koppelman (1999), Roorda (2005), Habib (2007), Bowman (2009), Feil (2010), Pinjari and Bhat (2011), Miller (2014) and Rasouli and Timmermans (2014).

All models of activity choice are built on the assumption that demand for traveling stems from a demand for activity participation, and understanding activity patterns is important for demand modeling. The goal of these models is to improve the traditional 4-step model by considering demand for activities instead of demand for trips (Bowman; 2009; Pinjari and Bhat; 2011). They model the interactions between time and space, trying to limit the independence assumption of the different elements of the choice (Rasouli and Timmermans; 2014).

This field of research faces methodological and technical challenges such as

- modeling interactions between different activities in the sequence (e.g., if there was a drop-off, there may be a pick-up later on; if you go shopping in the afternoon, you may not need to do it the evening; if you have dinner at 6pm, you don't need to do it again at 8pm);
- modeling social interactions: coupling constraints for household activities (e.g., Ho and Mulley; 2013; Gupta and Vovsha; 2013) or group activities from the social network;
- dynamic modeling demand across several days (e.g., Nijland et al.; 2014); or
- modeling matching of demand and supply, i.e., congestion, as in Bradley et al. (2010); Ouyang et al. (2011); Habib et al. (2013).

The activity-based approach is traditionally used to evaluate transport policies. The application of models of activity choice is now extending to new domains (Rasouli and Timmermans; 2014). They are used to quantify emissions (e.g., Shiftan et al.; 2015), health-related indicators (e.g., Perchoux et al.; 2013), carpooling (Galland et al.; 2014) and well-being (Abou-Zeid and Ben-Akiva; 2012; De Vos et al.; 2013).

Activity-based models can also be applied to pedestrians. The impact of timetables and platform allocations have been identified as a major challenge in pedestrian facilities such as train stations (Daamen; 2004, ch. 2) but is not studied in her thesis. Reviews about activity

choice for pedestrians can be found in Timmermans et al. (1992) and Bierlaire and Robin (2009). Recently, Liu (2013) developed an activity-based travel demand model in the context of an airport, focusing on activity scheduling, destination and route choice and rescheduling models. It is based on revealed and stated preference survey data. The revealed preference survey was sent to faculty and staff from the author's university. They were asked to describe the activities they performed and the activity with the longest duration the last time they visited an airport in the last 12 months. 359 responses were used for estimating the model. This thesis particularly focuses on congestion and consequent rescheduling. For Liu (2013), the home-based structure of urban activity behavior is replaced in airports by three structuring events: check-in, security check and boarding. Supposedly, in pedestrian context, the level of service (congestion, queues and flight schedules) is more important than people's characteristics, compared to activity choice in urban context. About the model, a nested logit is used. Each choice of activity is considered independently from other activities from a same individual. The nesting structure does not reflect any intuitive behavior.

In the following review of activity-based models, we are interested in how the time representation impacts the modeling of choice set generation in activity-based modeling. Different representations of time and activity types drive different strategies to manage the very large number of alternatives. In this review of activity-based travel demand modeling, we focus on the time representation (Section 2.3.3), the choice set generation (Section 2.3.4), the formulation of the utility function (Section 2.3.5) and the correlation between activity patterns (Section 2.3.6).

2.3.3 Time representation in activity modeling

Time is the most important dimension in modeling activity-travel behavior (Yamamoto and Kitamura; 1999). Its adequate representation is a prerequisite for accurate forecasting (Hägerstraand; 1970; Pinjari and Bhat; 2011). Compared to trip-based models, scheduling models see time not as a cost, but as a resource being used (Bhat; 2005). Time can be represented as continuous, decomposed in tours from home, or as the chronological activity episodes (e.g., Li and Lee; 2014).

An example of continuous time is the dynamic discrete-continuous choice model by Habib (2011). In this model, the decision maker sequentially chooses an activity type, and then its duration, constrained by a given time budget and the expected utility for the remaining time. This model is myopic, in the sense that it assumes that decision makers sequentially choose activities, without planning later activities. The concept of composite activity covering the rest of the day does not allow to model the choice of an overall pattern of activities with planning behavior. The multiple discrete-continuous extreme value models (Bhat; 2005; Pinjari and Bhat; 2011; Wang and Li; 2011) represent time as continuous and as a constraint (time budget). The decision maker chooses several activity types and allocates to each of them a corresponding time-use, so that it sums to the total time budget. The order in which it is

consumed is not modeled.

In the activity-schedule approach (Ben-Akiva et al.; 1996; Bowman; 1998; Bowman and Ben-Akiva; 2001; Bradley et al.; 2010; Shiftan and Ben-Akiva; 2011; Gupta and Vovsha; 2013), the fundamental unit of time is a tour. A tour is a way to decompose the available time in a day in manageable units with duration. The behavior is modeled as two sequential decisions, the activity pattern and the tours. Only the tour models include information about timing, through (1) the tour time of day choice model, modeling the tour primary destination arrival and departure times, and (2) the trip departure time choice model, modeling some intermediate stop arrival and departure times for trips on the tour (Childress; 2010; Abou-Zeid and Ben-Akiva; 2012). Tours, primary destination of the tours and subtrips on the tours are sequentially modeling the timing of activities in order to simplify the model and reduce the size of the choice set.

The time representation in activity modeling is generally continuous or tour-based. In continuous time, the time representation is close to reality but misses an overall pattern choice (choice of time expenditure per activity type without order in activity episodes in e.g., Pinjari and Bhat; 2011 or choice of time expenditure per activity episode without pattern utility in Habib; 2011). With tours, the choice of pattern is explicit but independent of the timing decision (start time and duration). It assumes a primary activity of the tour and a main mode (e.g., Shiftan and Ben-Akiva; 2011).

2.3.4 Choice set generation

Choice set generation is the process of defining the considered alternatives in an individual decision making. Assumptions must be made about the availability of the different options and the decision maker's awareness of them. Availability or awareness of the alternatives can be deterministically or stochastically defined (Ben-Akiva and Bierlaire; 2003).

Choice set generation in route choice

In the route choice context, the number of paths connecting an origin and a destination is very large and cannot be enumerated in practice. The universal choice set, containing all possible routes, cannot be used. There are two ways of dealing with it: selecting a choice set that only contains the paths considered by the decision maker (consideration choice set), or sampling a subset of paths large enough to be confident that it contains all important paths for the decisions maker (importance sampling). The consideration choice set is supposed to be consistent with behavior but is very often not available and too small for estimation, while the importance sampling is not very realistic behaviorally but is statistically more efficient.

Van Nes et al. (2008) propose a classification of different choice sets, by order of inclusion of alternatives: the chosen alternative, the considered alternatives, the reported alternatives (by the decision makers as considered), the feasible alternatives (i.e., available), the logical

alternatives (i.e., no loop) and the existing alternatives (i.e., the universal choice set). They compare choice sets made of reported alternatives by the respondents and choice sets made of feasible alternatives defined by a set of constraints. Having access to reported alternatives is difficult, and even impossible when using localization data from smartphones or antennas. Moreover, when data are available, the size of the choice set is too small for model estimation (average size of 2.8 in Van Nes et al.; 2008).

Consideration choice set Depending on the data collection technique, the consideration choice set can be explicitly asked in a survey. This is very often not possible, in particular when using different traces (GPS, WiFi, or other tracking systems). In these cases, the consideration must be modeled and a choice set generation algorithm is defined. It can be seen as a pre-choice before the actual choice. These models are not based on data but on assumptions about how people choose the paths they evaluate.

Repeated shortest path search

These approaches assume that people consider only the shortest paths as possible alternatives. This is less restrictive than it might appear, by using a generalized cost. The repeated shortest path approaches assume that the consideration set is made of a large enough number of shortest paths. These approaches generate very similar paths. In order to represent the heterogeneity of all paths and the variety of choices, van der Zijpp and Fiorenzo Catalano (2005) propose to remove unrealistic paths. Instead of generating a large number of shortest paths and removing the irrelevant ones, they propose algorithms for the constrained shortest path problem, directly generating feasible shortest paths. Constraints are supposed to express relevance, such as attractivity, circuitousness, non-overlapping or detour.

Constrained enumeration

These approaches assume that people do not consider some alternatives due to constraints. They generate all possible alternatives satisfying these constraints using branch-and-bound. Prato and Bekhor (2006) describe the construction of a connection tree between the origin and the destination. It is depth-first built and the branching rule is based on logical constraints: shortest path constraints with tolerance for going backwards or for longer travel times, avoiding detours, loops, overlaps and left turns. Parameters for these constraints are hand-tuned in order to reach behavioral consistency. Consistency is defined as heterogeneity and realism, i.e., ability to reproduce actual chosen routes. Prato and Bekhor (2006) collect 236 actual chosen routes to go to work using a web-based survey. 90 % of the respondents also reported the alternative routes considered for reaching the workplace, corresponding to 339 possible alternatives. Exploiting the road network of the city of Turin, Italy, 91.1% of chosen paths and 82.6% of reported paths are reproduced .

Importance sampling Flötteröd and Bierlaire (2013) propose a Metropolis-Hastings algo-

rithm for the sampling of paths. It creates a Markov chain sampling paths according to an arbitrary distribution, without enumerating all possible paths.

The goal consists in drawing a state $i \in S$ with probability $\frac{b(i)}{\sum_{i \in S} b(i)}$, where $b(i)$ is the target weight of state i. With this approach, the probability does not need to be computed and only the target weight $b(i)$ is needed. The target weights are defined as

$$b(i) = \frac{e^{-\mu\delta(\Gamma)}}{|\Gamma|(|\Gamma|-1)(|\Gamma|-2)/6} \tag{2.2}$$

where δ is the cost function of the path Γ, μ a scale factor and $|\Gamma|$ the number of nodes in path Γ. The denominator in the definition of $b(i)$ is justified by the state variable definition: the state variable i is defined as a tuple (Γ, a, b, c), with a, b and c nodes on path Γ, and consequently a path Γ corresponds to $|\Gamma|(|\Gamma|-1)(|\Gamma|-2)/6$ state variables i.

In the Metropolis-Hastings algorithm, the probability to move from state i to state j is needed. The transition matrix Q defines the proposal distribution and can be decomposed in two main operations (see Fig. 4.4). One operation, "splicing", randomly draws a node as a replacement of b and connect a and c through this new node with shortest paths according to the proposal distribution. The other operation, "shuffling", redistribute a, b and c along Γ. Flötteröd and Bierlaire (2013) propose to draw the new node with a logit distribution using shortest path length through this node, in order to drive the process toward short paths. To guarantee scale-invariance with respect to path cost, they suggest the use of a scale parameter $\mu = \frac{\ln 2}{(\zeta-1)\delta_{SP}}$. In this way, the probability of choosing a path of cost $\zeta\delta_{SP}$ is twice less than the shortest path (with cost δ_{SP}). For the second operation, "shuffling". they propose a uniform ascending choice of a, b and c.

These techniques allow to sample paths from a large network according to any sampling probabilities. The sampling probabilities do not need to be defined by link, but can be defined directly for the whole path. Importance sampling probabilities are crucial for an explicit correction in the discrete choice model.

Chen (2013) (Ch. 5) uses the Metropolis-Hastings path sampling technique by Flötteröd and Bierlaire (2013) for a route choice model estimated from GPS data. The weight function is composed of the path's length and of the frequency of observation of the given path. This "observation score" represents the inclusion of observed GPS data in the sampling process in order to include more relevant observations. This algorithm reduces the needed choice set size.

Choice set generation in activity choice

Choice set specification has received attention in different fields such as route choice or residential location choice (Ben-Akiva and Boccara; 1995; Swait; 2001; Başar and Bhat; 2004; de Lapparent; 2009; Rasouli et al.; 2013). A general review about choice set generation in

spatial context can be found in Pagliara and Timmermans (2009).

Activity scheduling consists of two steps: generating the choice set and making a choice among this set (Liao et al.; 2013). The definition of the choice set for each decision is a major weakness of the approach (Kang and Recker; 2013). The choice set generation in the context of activity modeling is complex and fundamental to have unbiased estimates for the parameters of the model, but data about the actual choice-set are usually missing. Moreover, the choice set containing all combinations of activity episodes is large (Bowman; 1998, p.74).

Most research in the field is applied to simpler problems. The size of the choice set is reduced through limitations in what enters the choice set and different assumptions related to the representation of time.

For each activity type $k \in \{1, ..., K\}$ (e.g., shopping), the activity-schedule approach considers two choices:

1. Is there a home-based tour with this activity type as the primary destination?
2. Are there secondary stops on this tour?

Thus, at a maximum, the choice set for the activity pattern consists of 2^{2K} alternatives (2^K possible tours with a primary destination being one of K activity types and, for each primary destination, 2^K tours with secondary stops being one of K activity types, see Abou-Zeid and Ben-Akiva; 2012 for an example). It is further reduced using logical rules, such as the impossibility to have secondary stops if no corresponding primary tour is chosen. Different applications of the activity-schedule approach use different sizes of the choice set in the activity pattern model (Shiftan and Ben-Akiva; 2011). Nevertheless, they all use the home-based tour structure to reach a manageable size for the activity pattern choice set (48 elements in the San Francisco model, 114 in the Portland model (Shiftan and Ben-Akiva; 2011)). The timing of the tours and stops is estimated in submodels.

In the case of activity sequences (Li and Lee; 2014), the choice set contains K^M sequences, assuming a maximum number M of possible activities in the day. Timing is not considered and thus it decreases the size of the choice set. Some models consider a single activity pattern, typically home-work-home, with composite activity episodes for before and after work (Ettema et al.; 2007; Jenelius et al.; 2011). In these cases, the choice is only about timing. In the multiple discrete-continuous extreme value models, the choice set only contains K alternatives, one for each activity type, with decreasing utility with time.

As a general strategy, activity type and scheduling are often not considered simultaneously. By considering tours or sequential order, without timing, the size of the choice set decreases. When considering simultaneously activity type and scheduling, assuming that the period of interest contains T time units, the number of alternatives would be K^T and the choice set size explodes. Two answers are proposed in the literature: (1) Considering time as a continu-

ous variable and using a multiple discrete-continuous approach (Bhat; 2005) or a dynamic discrete-continuous approach (Habib; 2011), or (2) Using the universal choice set containing all possible combinations of time units and activity types and performing importance sampling, as mentioned in Flötteröd and Bierlaire (2013). In this case, Lemp and Kockelman (2012) propose a first estimation using simple random sampling (SRS) and a second estimation drawing alternatives in proportion to the choice-probability estimates.

In the pedestrian context, activity choice is considered as an important level of pedestrian behavior in Daamen (2004, ch. 2) but is not dealt with because of the challenge of the choice set generation (Daamen; 2004, ch. 3). In an airport context, Liu, Usher and Strawderman (2014) consider three time units (between entering and check-in, between check-in and security, between security and boarding) and for each time unit the choice to perform or not five activity types, resulting in 15 alternatives. They recognize the need of modeling the various sequences for a future study.

2.3.5 Formulation of utility function for activity sequences

Modeling the activity pattern formation requires a proper definition of the utility of an activity pattern. Bhat (2005) considers the overall utility on an individual as the sum of the utilities of each activity episode. This utility must reflect the time-of-day preferences, the fatigue effects and the scheduling constraints (Ettema et al.; 2007). The activity-schedule approach also defines the primary activity as the most important activity of the day (Bowman and Ben-Akiva; 2001).

Time of day preference

The time-of-day element of the utility represents the variation in gain from performing the activity at different periods of time ("when"-dimension). Ettema and Timmermans (2003); Joh et al. (2004); Jenelius et al. (2011); Fu and Lam (2014) assume that the marginal utility follows a unimodal function, increasing first for a warming up phase and decreasing after reaching a saturation point. Ettema et al. (2004) propose to use a Cauchy distribution to express the marginal utility.

Satiation effect

The utility of an activity episode increases with time, while marginal utility of activity participation decreases (Yamamoto and Kitamura; 1999; Ettema et al.; 2007; Pinjari and Bhat; 2010; Habib; 2011). Satiation[2] expresses a fatigue effect ("how long"-dimension). In Habib (2011),

[2] Researchers in the transportation research community use "satiation" in order to express the decreasing marginal returns in activity modeling literature. We comply with this usage. Still, it must not be confused with the non-satiation assumption of preferences in standard microeconomics, stating that more is always better, or more precisely that, regardless of the individual's comsumption, an arbitrarily small quantity of a good is generating

the utility of time expenditure is multiplied by $\frac{1}{\alpha}(t^{\alpha}-1)$, with α the satiation parameter. The role of the satiation parameter α is to reduce the marginal utility with increasing duration. When $\alpha = 1$, there is no satiation and utility linearly increases with time expenditure. Satiation appears when $\alpha < 0$ (Bhat; 2008). This is why Habib (2011) specifies $\alpha = 1 - \exp(-\tau y)$, where y is a vector of variables and τ is the corresponding parameter vector. In practice, α depends on a constant per activity type and on the time-of-day. The effect of these two variables on satiation is evaluated by estimating τ.

Ettema et al. (2007) assume that a part of the utility for an engagement in an activity depends on the duration. This duration dependent utility follows a logarithmic function for the fatigue effect, $\eta_k \ln(t)$. η_k is specific to the activity type. The corresponding marginal utility is $\frac{\eta_k}{t}$. When the duration t increases, the duration dependent utility $\eta_k \ln(t)$ increases and its marginal counterpart $\frac{\eta_k}{t}$ decreases.

Schedule constraints

Time of day preference and fatigue effect have been used in several papers (Ettema and Timmermans; 2003; Jenelius et al.; 2011). Ettema et al. (2007) add schedule constraints in the utility function, inspired by Small (1982). First introduced by Vickrey (1969), scheduling costs explain the choice of departure time (e.g., Arnott et al.; 1990, 1993). Small (1982) defines the utility of trip departure time as a function of travel time and schedule delay:

$$V = \alpha_{tt} tt + \gamma_e \text{SDE} + \gamma_l \text{SDL} \tag{2.3}$$

where tt is the travel time, SDE is the early schedule delay and SDL is the late schedule delay. Schedule delays are defined as the difference between the preferred arrival time t^* and the actual arrival time t^a:

$$\text{SDE} = \max(t^* - t^a, 0) \tag{2.4}$$

$$\text{SDL} = \max(t^a - t^*, 0) \tag{2.5}$$

Schedule delay approach introduces constraints in the schedule, such as a train departure time, preferred time to start working, or beginning of courses. They are a mathematical expression of coupling constraints as defined by Hägerstraand (1970). Ettema et al. (2007) and Hess et al. (2007) include schedule delay in models of choice of activity patterns. This approach has some limitations. It assumes a linear, yet asymmetric, effect of being early or late. More recent works propose non-constant marginal utilities for the scheduling preferences (Tseng and Verhoef; 2008). In order to implement schedule constraints, anchor points t^* need to be known.

positive utility. For a proper use of "satiation" in transportation research, see Kockelman (1998, 2001).

2.3.6 Correlation between activity patterns

The correlation among alternatives (i.e., paths) is a well-known issue in route choice modeling when using random utility models. The similarity between two paths is usually measured as physical overlap (Vovsha and Bekhor; 1998). Frejinger and Bierlaire (2007) decompose the strategies to address this issue in two categories of models: deterministic correction (C-Logit models byCascetta et al.; 1996, Path Size Logit by Ben-Akiva and Bierlaire; 1999) and explicit modeling of correlation in the error term (Cross Nested Logit by Lai and Bierlaire; 2014, Error Component models by Bolduc and Ben-Akiva; 1991). The first category is the most frequent and the simpler to compute.

The Path Size Logit consists of including an attribute, called Path Size (PS), in the deterministic part of the utility, in order to correct for overlapping paths. It is derived from aggregation of alternatives (Ben-Akiva and Lerman; 1985, ch. 9), where the elemental alternatives are the paths and the aggregate alternatives are the links. In the route choice context, the size of the aggregate alternatives, i.e., a link, equals the number of paths using the link (see Frejinger and Bierlaire (2007) for a detailed development).

In activity choice context, utility is related to time of day, satiation and schedule delay (Section 2.3.5). All these parameters are defined for a given activity type (e.g., shopping). In the Activity Schedule approach, Bowman (1998, p.25) mentions that patterns sharing primary purpose are probably correlated and let it as a future research. Primary activity is defined as the most important activity of the day, either by asking the respondents in a survey (Antonisse et al.; 1986) or by counting the number of activity episodes in a tour and using deterministic priority rules from other studies when not available from the survey (Bowman and Ben-Akiva; 2001). Note that when deterministic rules cannot discriminate between different activity purposes, the activity of longer duration is defined primary.

As a general conclusion, this literature review for models of activity choice shows the need for modeling an explicit pattern utility with timing dimension (start time and duration). Considering simultaneously the order, timing and duration of activity-episode patterns generates very large choice sets. Large choice sets have been managed with discrete-continuous approaches. Importance sampling techniques have been recently improved and are an interesting approach, but have not been implemented yet. Since 1998, correlation between activity patterns have been detected as an issue but no answer has been given to it to our knowledge.

2.4 Location choice

Location choice models are common in studies of urban transportation policies and planning. Ben-Akiva and Lerman (1985) mention three of them, for the Paris region and Maceio, Brazil. Very often, such models are applied to the choice of location for grocery shopping (Timmermans; 1996; Dellaert et al.; 1998; Fox et al.; 2004; Scott and He; 2012). Location choice models are applied to several other different contexts, such as the choice of a departure airport

(Furuichi and Koppelman; 1994), the choice of a hospital for patients by general practitioner (primary care physicians) (Whynes et al.; 1996), the choice of touristic destinations (Woodside and Lysonski; 1989; Um and Crompton; 1990; Eymann and Ronning; 1997; Oppermann; 2000; Seddighi and Theocharous; 2002; Bigano et al.; 2006; Chi and Qu; 2008; Gössling et al.; 2012; Yang et al.; 2013) and in particular recreational outdoor facilities (Fesenmaier; 1988; Scarpa and Thiene; 2005; Thiene and Scarpa; 2009), the choice of migrants (Fotheringham; 1986) and the optimal allocation of charging stations for electric vehicles (He et al.; 2013).

Regarding pedestrians, Borgers and Timmermans (1986a) successively model the destination choice to predict the total demand for shops in a shopping street in a city center. Timmermans et al. (1992) review the existing literature in 1992. Eash (1999) has developed models for non motorized destination choice, with application to the Chicago Area. Zhu and Timmermans (2011) propose heuristics rules including principles of bounding rationality and compare them to discrete choice models. The models are validated on the same sample used for estimation and no cross validation is performed. Ton (2014) studies route and location choice in train stations based on tracking and counting data. Counting data come from infrared scanners and tracking data come from WiFi and Bluetooth scanners. Counting data allow to apply the model to pedestrians without smartphones. The choice is between locations for a given activity type. Kalakou et al. (2014) apply a similar approach for location choice for a given activity type (which coffee shop knowing that the individual is visiting one) in an airport.

2.4.1 Attributes of the choice for a location

The main attributes in location choices in urban context are travel time, travel cost and distance (Cambridge Systematics Europe; 1984; Ben-Akiva and Lerman; 1985; Whynes et al.; 1996). Other variables are used: type of neighborhood, number of different services or speciality stores (banks, post offices, medical facilities, offices, shops, foodcourts, cinemas, etc.), parking facilities (park-seek time, parking cost), number of retail employees (Cambridge Systematics Europe; 1984; Ben-Akiva and Lerman; 1985; Zhu et al.; 2006; Shobeirinejad et al.; 2013) or symbolic acts (support of community charities, front-door greeters, patriotic displays) (Arnold et al.; 1996). Another typical attribute is the *size* in the context of aggregation of alternatives (see Section 2.4.2). It represents the number of elemental alternatives in the considered aggregate alternatives (subsets of the choice set). The interpretation of this attribute is complicated, since it absorbs both the preference for a large set of destinations compared to a small one and the correlation between destinations in the set. The expected sign is opposite in the two situations (Frejinger and Bierlaire; 2007). In shop patronage, the main attributes are the retail floor space, the accessibility and the price (Arnold et al.; 1983; Scott and He; 2012).

In the pedestrian context, the main attributes of location choice are the attractiveness of the location and travel time. More specifically, models include floor space (Borgers and Timmermans; 1986a), pedestrian environment in neighborhood, employment (Eash; 1999) as measures of attractiveness and distance as an approximation of travel time (Borgers and

Timmermans; 1986a; Ton; 2014). Kalakou et al. (2014) include space syntax in the specification of the utility through "integration", i.e., a measure of accessibility.

2.4.2 Location choice models

In an urban context, models are often based on tours, characterized by a travel mode and a destination. Models of joint choice of travel mode and destination aggregate destinations into zones (Cambridge Systematics Europe; 1984; Ben-Akiva and Lerman; 1985). Stratified importance sampling is used, dividing the destination choice set into non-overlapping strata based on the origin zone. In the Paris Region example, this procedure decreases the choice set from 595 destinations × 4 travel modes to 7 sampled alternatives for each trip (Cambridge Systematics Europe; 1984; Ben-Akiva and Lerman; 1985). In a pedestrian context, the choice set is often smaller, due to the smaller study area (e.g., Ton; 2014; Kalakou et al.; 2014, with 2 to 4 alternatives). Most destination choice models are logit models (Arnold et al.; 1983; Zhu et al.; 2006; Scott and He; 2012; Kalakou et al.; 2014; Ton; 2014). Probit models have been used (e.g., Whynes et al.; 1996).

Panel data are common in transportation research (Golob et al.; 1997) and habits are often observed in travel behavior (Gärling and Axhausen; 2003), in particular in route choice (Aarts and Dijksterhuis; 2000; Bamberg et al.; 2003; Thøgersen; 2006; Eriksson et al.; 2008; Verplanken et al.; 2008; Gardner; 2009; Schwanen et al.; 2012) and in car ownership (Jong et al.; 2004). In Markov models of destination choices, transition matrix represents the probability of choosing a destination given the choice of destination at the previous stop. Markov models are criticized for being descriptive, replicating the data, and not being sensitive to behavioral changes (Kitamura; 1990; Timmermans et al.; 1992). Dynamic models using panel data increase statistical efficiency, improve predictions and allow to study behavioral dynamics (Kitamura; 1990). In 1990, Kitamura (1990) considered the inclusion of lag terms in discrete choice models not well advanced. Unresolved issues in the estimation of dynamic models using panel data included the representation of the initial conditions and the correlated error term in dynamic models. In 2001, McFadden (2001) highlighted the importance of panel data in discrete choice models (Carrel et al.; 2015). Yang et al. (2013) model the choice of a second touristic destination after visiting a first one using a nested logit. The panel nature of the data is not explicitly taken into account in their model, similarly to Wu (2012, ch. 5.2), since the previous destination characteristics are not included in their model. For pedestrian destination models, Timmermans et al. (1992) mention in their review the "issue of whether a pedestrian tends to always buy certain items in the same store", i.e., the question of loyalty, as future research.

Few authors explicitly include lagged variables in location models. In tourism literature, Grigolon et al. (2014) include the previous vacation length choice in the choice of the current vacation length. They compare a logit, a mixed logit and a dynamic mixed logit and show that the dynamic mixed logit is the best in estimation and forecasting. In their dynamic mixed logit, by assuming that the error term is independent of the variables (i.e., exogenous), and in

particular independent of the lagged variable, they assume that unobserved attributes do not persist over time for a given individual. This can lead to bias in the estimation of the model, in particular when the choice of the vacation length of a stay is influenced by variables not included in their model. In the choice of a shop in a pedestrian street, Zhu et al. (2006) also face serial correlation and mention independence issues as a technical challenge for future research.

In light of the literature review in Section 2.2.1 and 2.4, we emphasize that location-aware technologies allow to collect panel data in the long term. These data must be used in location choice models and lagged variables must be included in the utility function of locations. This leads to bias in the estimation of the model and serial correlation. Wooldridge (2005) proposes a solution to the problem of serial correlation (see Section 5.2 for details). It is mostly applied to binary probit (Arulampalam and Stewart; 2009). Our contribution in Ch. 5 develops a location choice model using panel data from localization-aware technologies. We include a lagged variable and Wooldrige's correction for endogeneity. To our knowledge, this correction has never been applied to dynamic location choice. We apply it to a logit model with 21 locations in the choice set.

Preprocessing Part I

3 Detecting activity-episode sequences

In collaboration with Bilal Farooq

3.1 Introduction

Individual mobility traces are becoming available from pervasive systems, such as cellular networks or WiFi hotspots (see Section 2.1.1). In many cases, cost and privacy issues prohibit from installing high precision sensors such as cameras covering an entire pedestrian infrastructure. The large size of pedestrian facilities, such as an airport or a railway station, implies either precise sensors with incomplete coverage (e.g., cameras or bluetooth sensors in intersections), or full coverage with imprecise long range sensors (e.g., cellular network data, traces from WiFi infrastructures). As a result, localization data are either sparse, fuzzy, or both. We propose a methodology exploiting sparse data with an explicit modeling of the imprecision in the measure, and using prior knowledge of the infrastructure. It provides mobility patterns semantic, such as stop locations and purposes and start and end times. The methodology recovers mobility patterns that generated the localization data, similarly to reverse geocoding, i.e., recovering a postal address from x-y coordinates.

Section 3.2 describes the necessary data for detecting pedestrians, while Section 3.3 describes the methodology to merge these data. A case study on the Ecole Polytechnique Fédérale de Lausanne (EPFL) campus is described in Section 3.4, with results of this case study, together with validation and sensitivity analysis. Finally, we conclude and discuss future work in Section 3.5.

3.2 Data requirement

We assume two kinds of data sources before applying the detection methodology: network traces and semantically-enriched routing graph with which we can associate the network traces.

3.2.1 Network traces

An input of the probabilistic detection method consists of timestamps and localization data coming from network traces: WiFi traces, GSM traces, Bluetooth tracking or RFID localization. We define a measurement as $\hat{m} = (\hat{x}, \hat{t})$, where $\hat{x} \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}$ is the position of the measurement (x-y coordinates in a coordinate system, and floor or altitude in a multi-floor environment) and $\hat{t}$ the measurement timestamp. The nature of the measurement depends on the data collection technique. With data from access points (APs) (WiFi APs, cell towers, ...), the measurement $\hat{x}$ corresponds to the position of an AP; in multilateration data, the measurement $\hat{x}$ is anywhere in space, and not related to AP locations. For a given individual n, we assume a chronologically ordered sequence $(\hat{m}_{1,n},,...,\hat{m}_{j_n,n}...,\hat{m}_{J_n,n})$, which is abbreviated as $\hat{m}_{1:J_n,n}$, where J_n is the total number of measurements. The measurement timestamp $\hat{t}$ is continuous.

Accuracy ξ is also needed for each measurement $\hat{x}$. It is defined as the distribution of the Euclidean distance between the location estimate $\hat{x}$ and the actual location $\mathring{x}$, $\hat{x} = \mathring{x} + \xi$. It can either be constructed by the information provided by the localization tool (e.g. level of confidence, attenuation rate, etc.) or by the analyst. In the second case, one has to design experiments and calibrate the error distribution based on already known locations in the area covered by the antennas.

Different levels of anonymity are possible with these data. Originally, a unique identifier per device is collected (e.g., the MAC address for a smartphone using WiFi). This unique identifier may be processed in two different ways. First, it may be associated with a username through identification in the system and thus to the identity and socioeconomic information if available, such as gender, age or income. Second, it may be anonymized to guarantee anonymity. The data anonymization can be total or keep some socioeconomic information. This way, i can

1. correspond to a unique anonymous ID, just for tracking individual traces;
2. be associated with some socioeconomic characteristics without the identity of the user, or
3. correspond to a personal identifier.

All options are technically feasible. EPFL ethics committee recommends to remove the personal identifier when sharing the data with other researchers (allowing options 1 or 2 and banning option 3 for transmission).

3.2.2 Pedestrian semantically-enriched routing graph

The following detection methodology needs a *semantically-enriched routing graph* (*SERG*). We define *SERG* as a set of nodes $\mathcal{N}$ and a set of edges $\mathcal{E}$. *SERG* allows for routing individuals from origins to destinations through an optimal path, and contains information about points

of interest (e.g., the name of the room or the type of the room in a pedestrian context, or shopping facilities in an urban context). In order to link localization measurements $\hat{x}$ to the graph, each node in $\mathcal{N}$ must be associated with a coordinate system.

In a centerline approach as defined in Goetz and Zipf (2011) for a corridor in an indoor context, some nodes correspond to intersections and not to possible destinations. Nodes are defined as destinations if they correspond to a room, a shop or a restaurant, i.e. if they are points of interest (POI) in the pedestrian infrastructure. POI is a subset of $\mathcal{N}$.

Formally, $SERG := (\mathcal{N}, \mathcal{E}, \mathcal{L}, f, g, POI)$, where $\mathcal{L}$ is a set of relevant labels for rooms, restaurants, shops, etc., $f : \mathcal{N} \rightarrow \mathcal{L}$ is the labeling function, and $g : \mathcal{N} \rightarrow \mathbb{R} \times \mathbb{R} \times \mathbb{R}$ a function associating nodes with coordinates in a coordinate system. $POI \subset \mathcal{N}$.

More information can be added to $SERG$. The path realism (both physically and behaviorally) can be improved by adding information to the graph and using a generalized cost for the shortest path algorithm. A solution to balance between the shortest path and the simplest path is to give each edge of the pedestrian network a weight. It represents the aversion to floor changes and less important walkways in a pedestrian context, or to left turns or traffic lights in an urban context. Goetz and Zipf (2011) propose a *weighted indoor routing graph*, which is an enriched version of $SERG$. Adding information to edges $\mathcal{E}$ allows for a more realistic shortest path algorithm. Adding information to nodes $\mathcal{N}$ gives the opportunity to associate other data, such as schedules, opening hours, or door access.

3.3 Methodology

We are proposing a modeling approach to extract the possible activity-episode sequences performed by pedestrians from digital traces in a communication networks. This Bayesian approach merges measured network traces (continuous in space) (Section 3.2.1) and pedestrian semantically-enriched routing graph (Section 3.2.2) to compute the likelihood that a given sequence of activity episodes (discrete in space) has actually generated the observed traces.

We define an activity episode $a = (x, t^-, t^+)$ as a POI where the user is spending time, where x is the episode location, t^- the episode start time, and t^+ the episode end time. The episode location x is a POI in SERG, $x \in POI$, and is labeled, $f(x) \in \mathcal{L}$. In the wireless localization literature, the episode location x is called a *symbolic location*, "a location in a natural-language way" (Liu et al.; 2007). t^- and t^+ are continuous random variables and define the time spent at destination, $t^+ - t^-$. We impose that $t^+ - t^- \geqslant T_{\min}$, a minimum threshold (typically 5 min in a pedestrian context, or 20 min in the same antenna location in an urban context in Bekhor et al.; 2013). The output of the probabilistic method consists of a set of L candidate activity-episode sequences $(a_1, ..., a_{\psi_n}, ..., a_{\Psi_n})$, which is abbreviated as $a_{1:\Psi_n}$, where Ψ_n is the total number of episodes. Ψ_n is individual specific and unknown to the analyst. In the following developments, both the number of measurements in the sequence J_n and the number of episodes Ψ_n are individual specific, but the n subscript is omitted to make the notation light. Each candidate

activity-episode sequence $a_{1:\Psi}$ is associated with the probability of being the actual one.

In the next section, we propose a probabilistic measurement model associated with an activity-episode sequence. Then, in Section 3.3.2, the generation process of candidate activity-episode sequences is described. Figure 3.1 shows the plate model (see Koller and Friedman; 2009, Section 6.4.1) of the link between the activity episodes and the measurements.

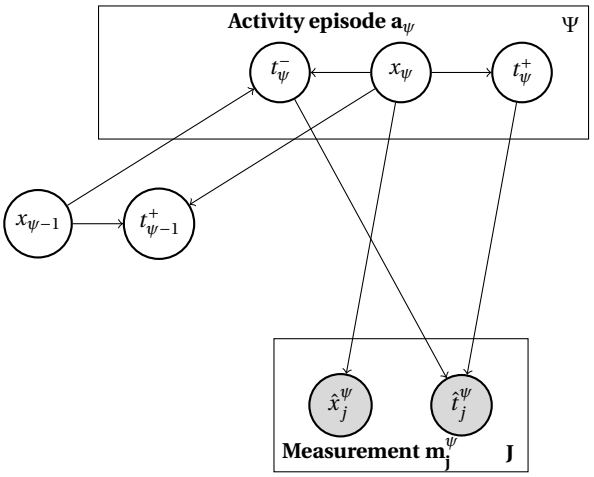


Figure 3.1 – Plate model for the probabilistic measurement model. It represents the generation process of network traces. While being in point of interest x_ψ between times t_ψ^- and t_ψ^+, users generate measurements $\hat{x}_j^\psi$ at time $\hat{t}_j^\psi$. Dark shaded nodes represent the observed variables. The arrows represent the dependencies between the variables. Boxes express the multiple iterations of the conceptual object: there are Ψ activity episodes a_ψ, and J measurements m_j^ψ in total.

3.3.1 Probabilistic measurement model: a Bayesian approach

A probability $P(a_{1:\Psi}|\hat{m}_{1:J})$ is associated with each activity-episode sequence $a_{1:\Psi}$. It takes into account the inaccuracy in the network traces based on the measurements and some prior knowledge about the potential activity-episode locations. The activity probability $P(a_{1:\Psi}|\hat{m}_{1:J})$ that $a_{1:\Psi}$ is the actual activity-episode sequence given the measurement $\hat{m}_{1:J}$ is decomposed as:

$$P(a_{1:\Psi}|\hat{m}_{1:J}) \quad \propto \quad P(\hat{m}_{1:J}|a_{1:\Psi}) \cdot P(a_{1:\Psi}) \tag{3.1}$$

where $P(\hat{m}_{1:J}|a_{1:\Psi})$ is the measurement likelihood and $P(a_{1:\Psi})$ is a prior knowledge about the activity episodes.

Measurement likelihood

For each activity-episode sequence $a_{1:\Psi}$, our goal is to compute the probability

$$P(\hat{m}_{1:J}|a_{1:\Psi}) \tag{3.2}$$

that the performed episodes generated the observed measurement sequence.

We assume that a measurement $\hat{m}_j$ always corresponds to an activity episode a_ψ. We denote $\hat{m}_j^\psi = (\hat{x}_j^\psi, \hat{t}_j^\psi)$ the measurement in $\hat{m}_{1:J}$ corresponding to $a_\psi = (x_\psi, t_\psi^-, t_\psi^+)$, i.e. when $t_\psi^- \leqslant \hat{t}^\psi \leqslant t_\psi^+$. As a result, $\hat{m}_{1:J} = \cup_\psi \hat{m}_{1:J}^\psi$. If a measurement is generated while walking, the model will consider it as a very short activity episode (that can be eliminated later).

If the device's owner is performing activity episode a, the probability that it will generate a measurement $\hat{m}$ is a function of the location of the episode location x, the measurement location $\hat{x}$ and the accuracy ξ of the measurement. Thus we can decompose Eq. 3.2 as:

$$\begin{aligned} P(\hat{m}_{1:J}|a_{1:\Psi}) &= \prod_{\psi=1}^{\Psi} P(\hat{m}_{1:J}^\psi|a_\psi) && (3.3)\\ &= \prod_{\psi=1}^{\Psi}\prod_{j=1}^{J} P(\hat{m}_j^\psi|a_\psi) && (3.4)\\ &= \prod_{\psi=1}^{\Psi}\prod_{j=1}^{J} P(\hat{x}_j^\psi|x_\psi). && (3.5) \end{aligned}$$

Equality in Eq. 3.3 assumes measurement independence between activities, i.e. measurement error in the sequence is only related to the corresponding activity episode in time. Equality in Eq. 3.4 assumes independence between measurements, i.e. error is the same for different measurements while in the same location x_ψ and time interval t_ψ^-, t_ψ^+. Equality in Eq. 3.5 assumes no measurement error in time, i.e. measurement error is only a localization error.

These three independence assumptions mean that the error in measurement, i.e., the probability of being away from the real location, depends only on distance between a measurement and the real location, and on accuracy ξ. It does not mean that measurements are independent, but conditionally independent, knowing the real location of the device. Autocorrelation of the error between different measurements (for a given location) is due to obstacles or other environmental clutter (i.e., *shadowing*). Autocorrelation is assumed to be taken into account by the data collection tool, when generating $\hat{x}$ (see Mailaender; 2011; Taylor; 2013 about autocorrelation between measurements during the data collection process).

Prior: potential attractivity measure

Introducing prior knowledge may be needed when localization is weak and the density of POI is high (this is particularly common in a pedestrian context). Moreover, the prior gives the possibility to add information from available data. In this section, we propose the first formal definition of attractivity for pedestrian infrastructures to our knowledge. It is built on existing literature for urban context and allows to coherently merge different data sources.

Space-time accessibility is a very common concept in land use planning (Miller; 2010). Even before the accessibility literature, the classical gravity-potential indicators were defining similar concepts of attraction. Stewart (1948) defines a "population potential" in a law of "demographic gravitation" by substituting the mass by the number of people in Newtonian gravitation. Lakshmanan and Hansen (1965) measure attractivity in a retail market potential model as shopping goods floor space, referring to it as "supply". Carrothers (1956) proposes a review of the gravity and potential concepts. He shows that it has been applied to very different contexts (shopping center locations, population and migration forecasting, allocation of land use) with the central concept of population masses, or potential.

Space-time accessibility measures availability of activities for individuals given temporal and spatial constraints. Several definitions have been proposed. Hansen (1959) defines accessibility as a "potential of opportunities for interaction":

$$acc_i = \sum_j \frac{S_j}{tt_{ij}^{\alpha}}$$

where acc_i is the accessibility of place i, S_j is a measure of the "size of the activity" at j, such as the number of jobs, the annual retail sales or the population in a residential area, and tt_{ij} represents the travel time between i and j. The α parameter, defining the weight of travel time in accessibility measure, is evaluated based on the urban growth, assuming it is directly proportional to accessibility. Weibull (1980) develops a rigorous axiomatic framework defining attraction-accessibility measure based on distance and attractivity (also called supply capacity). No clear definition of what exactly is attractivity is given but in an example about labor market, attractivity is defined as a function of the number of jobs and a demand potential for each zone (Weibull; 1976). He mentions that attractivity may be described as "offer", and gives the examples of places at day-nurseries and hospital beds.

A definition of accessibility merging the attractivity-accessibility measures (Weibull; 1980) and the constraints-oriented approach (see Section 2.3.1 and Hägerstraand; 1970) is proposed by Miller (2010). He emphasizes that a pure constraints-oriented approach gives each opportunity an equal weight, and, conversely, an attractivity-accessibility approach does not take into account temporal constraints. We propose to similarly define a potential attractivity measure by merging attractivity and time constraints for the pedestrian context.

Formally, we define the potential attractivity measure as a model of aggregated occupation

per point of interest (POI). The unit of attractivity is the number of persons. The potential attractivity measure $S_{x,n}(t^-, t^+)$ between a start time t^- and an end time t^+ for $x \in POI$ and individual n is time dependent and may differ across individuals. It depends on the instantaneous potential attractivity measure $S_{x,n}(t)$ at a given time t:

$$S_{x,n}(t^-, t^+) = \int_{t=t^-}^{t^+} S_{x,n}(t)dt. \tag{3.6}$$

In practice, time is discretized and the integration is replaced by a sum. The instantaneous potential attractivity measure depends on time-constraints and attractivity:

$$S_{x,n}(t) = \delta_{x,n}(t) \cdot att_n(x,t)$$

where $\delta_{x,n}(t)$ is a dummy variable for time-constraints such as schedules or opening hours, with value 1 if the POI is open or scheduled and 0 otherwise: opening hours of shops and restaurants, or timetables in the case of conferences, campuses, or public transport infrastructures. Timetables are individual-specific. Their availability depends on the level of anonymity for localization data (see Section 3.2.1).

Attractivity $att_n(x,t)$ is context-specific, as seen in the land use literature: number of jobs, annual retail sale, population per zone, places at day-nurseries, hospital beds. In the pedestrian facility context, data sources could be checkouts in supermarkets, metro card swapping data, concert tickets data, number of seats in a restaurant, number of employees per office, number of students in class, capacity of different zones in a stadium or a public transport infrastructure.

As a general guideline, the potential attractivity measure depends on the available information:

- If the attractivity is stable in time for a given POI x (e.g., an office on campus with a given number of employees and no explicit office hours), $\delta_{x,n}(t) = 1 \ \forall t$ and $S_{x,n}(t) = att(x)$;
- If the POI has opening hours (e.g. a shop on campus), $\delta_{x,n}(t) = 1$ for t in the opening hours and 0 otherwise, and consequently $S_{x,n}(t) = att(x)$ for t in the opening hours and 0 otherwise;
- If the POI has varying attractivity in time, $S_{x,n}(t) = att(x,t)$ with $att(x,t)$ being a step function (e.g. for classrooms with different numbers of students at different periods of the day) or any function representing the number of people in the POI per time (e.g., point-of-sale data for restaurants);
- If the attractivity varies for different people or categories of people, $S_{x,n}(t) = att_n(x,t)$ with different attractivity functions $att_n(x,t)$ for different individuals n (e.g., a classroom has different attractivities for employees and students on a campus).

With the potential attractivity measure properly defined, the prior can be built based on it.

The prior $P(a_{1:\Psi})$ is derived assuming that successive activity episode a_ψ are independent, for all ψ.

$$P(a_{1:\Psi}) = \prod_{\psi=1}^{\Psi} P(a_\psi) \tag{3.7}$$

$$= \prod_{\psi=1}^{\Psi} P(x_\psi, t_\psi^-, t_\psi^+) \tag{3.8}$$

$$= \prod_{\psi=1}^{\Psi} \frac{S_{x_\psi,n}(t_\psi^-, t_\psi^+)}{\sum_{x\in POI} S_{x,n}(t_\psi^-, t_\psi^+)} \tag{3.9}$$

The prior probability is proportional to the potential attractivity measure $S_{x_\psi,n}(t_\psi^-, t_\psi^+)$. We use this expression in the following case study. Another possible specification of the prior probability could be an exponentially increasing function $\exp(\mu S_{x_\psi,n}(t_\psi^-, t_\psi^+))$ of the potential attractivity measure $S_{x_\psi,n}(t_\psi^-, t_\psi^+)$:

$$P(a_{1:\Psi}) = \prod_{\psi=1}^{\Psi} \frac{\exp(\mu S_{x_\psi,n}(t_\psi^-, t_\psi^+))}{\sum_{x\in POI} \exp(\mu S_{x,n}(t_\psi^-, t_\psi^+))}$$

where $\mu \geqslant 0$ is a scale parameter. The scale parameter μ controls for the link between the potential attractivity measure $S_{x_\psi,n}(t_\psi^-, t_\psi^+)$ and the prior probability $P(a_{1:\Psi})$: if $\mu = 0$, the prior probability is uniform among all activity-episode sequences $a_{1:\Psi}$; if $\mu \to \infty$, the prior probability concentrates on the activity-episode sequence with the highest potential attractivity measure.

We define four specifications of the prior based on different assumptions on the available data: uniform, aggregate, disaggregate and diary.

Uniform If no information about the attractivity is available, a default assumption has to be used, and attractivity is fixed for all POI (similar to $\mu = 0$ in the exponentially increasing function specification). The corresponding prior is called "uniform".

Aggregate If information about attractivity and schedule is available, the quality of the prior depends on the level of anonymity of the network traces. Without personal information, a single aggregate prior is defined using the same time constraints for all individuals to define the potential attractivity measure for each location.

Disaggregate Disaggregate information about schedules may be available without knowing the identity of the individual: travelers (with trip schedules) and non-travelers in a transport hub, employees (with working hours) and visitors in a shop, etc. These information define one "disaggregate" prior per group. They come either directly from the

network traces (see Section 3.2.1) or from pattern recognition (e.g., in a railway station, individuals directly arriving on a platform are automatically travelers coming from a train).

Diary Finally, individual schedules can be used to define a "diary" prior. Due to respondent burden, individual schedules from activity schedule surveys are particularly difficult to collect (see, e.g., Chen et al.; 2010). This prior is important for establishing the consistency of our approach.

3.3.2 Generation of activity-episode sequences

The probabilistic measurement model presented in Section 3.3.1 computes the likelihood of a given activity-episode sequence $a_{1:\Psi}$. This section focuses on the generation of candidate activity-episode sequences. An algorithm is proposed to generate candidates from localization data and pedestrian semantically-enriched routing graph. At each new measurement $\hat{m}_j$ of $\hat{m}_{1:J}$, we build a list of candidates for the activity-episode location x and corresponding start and end times t^- and t^+.

Generating episode location

Inspired by the methodology developed by Bierlaire et al. (2013) for smartphone GPS data, we generate candidate episode locations for each measurement using the concept of *domain of data relevance* (DDR) originally introduced by Bierlaire and Frejinger (2008).

We define the DDR as a physical area in space where a measurement location is relevant. The definition of the area can be different depending on the precision of the measurement, i.e., the DDR depends on the type of data. In a pedestrian context, this area could even be in 3D, covering several floors. A point of interest $x \in POI$ is considered to be in the DDR of measurement location $\hat{x}$ if the probability $P(\hat{x}|x) \geqslant \theta$, with θ a given threshold. This probability is a function of the location x of the POI and the measurement location $\hat{x}$, similar to Eq. 3.5.

Using the domain of data relevance DDR_j for each measurement $\hat{m}_j$, we generate all possible activity episodes for each individual in DDR_j. Each point of interest in this domain of data relevance, $x_j \in POI \cap DDR_j$, represents a possible episode location. It is connected with all possible next episode locations contained in the domain of data relevance DDR_{j+1} of the following measurement in time, $\hat{m}_{j+1}$. A simple example with two DDRs containing respectively 3 and 2 points of interest is presented in Fig. 3.2. Note that two successive DDRs can overlap, resulting in common points of interest; the generated episode location can be the same, $x_j = x_{j+1}$, and corresponds to two measurements from the same place.

For a list $\hat{m}_1, \hat{m}_2, ..., \hat{m}_J$ of measurements associated with a given individual, the result of this process is a network structure with path in this network with length J. Each path in the network corresponds to a sequence $x_{1:J}$ of potential episode locations. This network is built

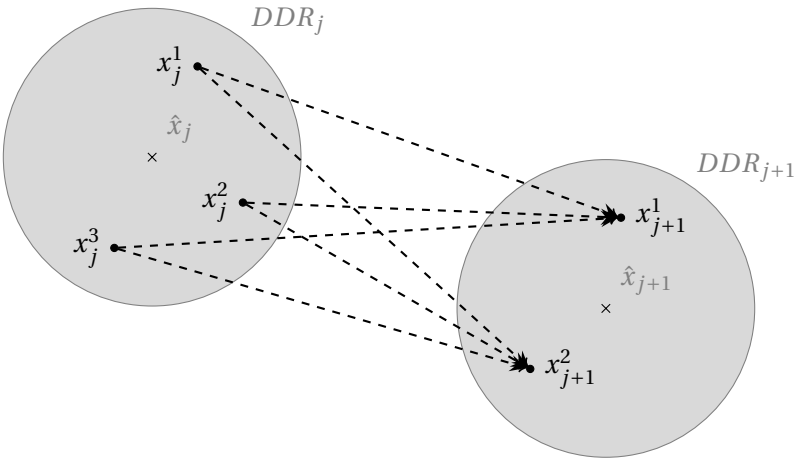


Figure 3.2 – The symbolic representation of two domains of data relevance DDR_j and DDR_{j+1} corresponding to measurements $\hat{m}_j$ and $\hat{m}_{j+1}$ following each other chronologically (in gray). In this simple example, we assume DDR_j contains 3 possible episode locations x_j^1, x_j^2, x_j^3, and DDR_{j+1} contains 2 possible episode locations x_{j+1}^1, x_{j+1}^2 ($x_j^1, x_j^2, x_j^3, x_{j+1}^1, x_{j+1}^2 \in POI$) (in black).

recursively. For each measurement $\hat{m}_j$, $j = 1, ..., J$ for a particular individual in chronological order, we consider all possible episode locations, i.e., all POI in $SERG$, in the domain of data relevance DDR_j. At each new measurement $\hat{m}_j$, the network structure of activity episodes is extended with all locations associated with $\hat{m}_j$ (Fig. 3.3).

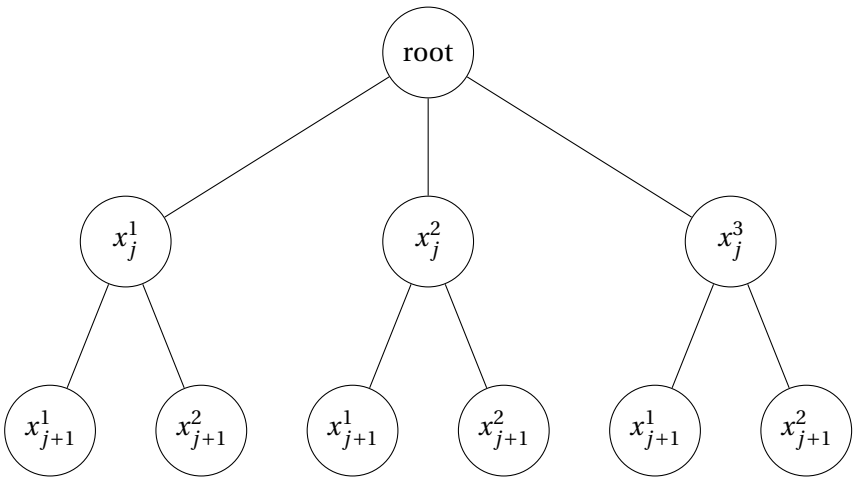


Figure 3.3 – A tree representation of the network corresponding to the two measurements in Fig. 3.2. Each path from the root to a leaf of the tree represents a possible activity-episode sequence.

If a measurement is imprecise and the corresponding DDR is huge (e.g., in an area with low WiFi coverage the size of the confidence interval can be of the order of magnitude of the whole pedestrian infrastructure), the prior is defining alone the most probable location as the point

of interest with the highest attractivity in the pedestrian infrastructure. An upper bound for the size of the DDR might be needed in these cases in order to limit the candidate activity-episode location, for efficiency. The activity episodes corresponding to these measurements containing little information will then be eliminated (see Section 3.3.3).

Generating episode start and end times

Once a sequence $x_{1:J}$ of potential episode locations is defined, the episode start and end times t^- and t^+ at these locations need to be generated.

Given two consecutive measurements $\hat{m}_j$ and $\hat{m}_{j+1}$ and their corresponding timestamps $\hat{t}_j$ and $\hat{t}_{j+1}$, a trip between the two generated positions x_j and x_{j+1} of the consecutive activity episodes a_j and a_{j+1} is assumed to take place. This trip defines both the end time t_j^+ from episode a_j and the start time t_{j+1}^- of episode a_{j+1}. The departure of the trip occurs after measurement $\hat{m}_j$ and before the latest possible departure time, i.e., the time that allows to reach the next episode location through the shortest path. Similarly, the arrival occurs before the next measurement $\hat{m}_{j+1}$ and after the trip from the episode location of the previous measurement (Fig. 3.4).

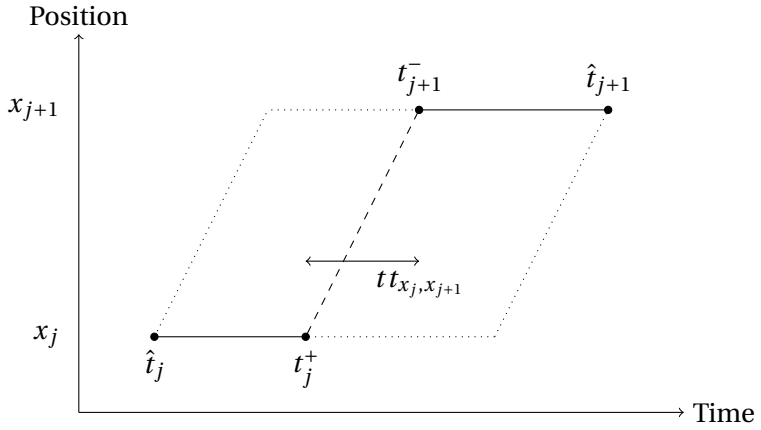


Figure 3.4 – Time-space representation of two consecutive activity episodes j and $j+1$.

The travel time used by a pedestrian to walk from x_j to x_{j+1} is approximated by the ratio between the shortest path distance between x_j and x_{j+1}, and the speed of 1.34 m/s (see Buchmüller and Weidmann; 2006). In this way, the episode end time t_j^+ is defined as $t_j^+ \in [\hat{t}_j, \max(\hat{t}_j, \hat{t}_{j+1} - tt_{x_j,x_{j+1}})]$ and the next episode start time t_{j+1}^- is defined as $t_{j+1}^- \in [\min(t_j^+ + tt_{x_j,x_{j+1}}, \hat{t}_{j+1}), \hat{t}_{j+1}]$. The maximum and the minimum in the bounds of the intervals manage the situation when $\hat{t}_j \geqslant \hat{t}_{j+1} - tt_{x_j,x_{j+1}}$ or $t_j^+ + tt_{x_j,x_{j+1}} \geqslant \hat{t}_{j+1}$. This may happen when the pedestrian was much faster than what we assume, or when a measurement was generated while walking (no stop, thus no time spent at this location).

No information is available about the exact time when the trip actually happens between the bounds for start and end times, and so a uniform distribution is used in all case studies in this

thesis. In the following mathematical developments, we assume that $\hat{t}_j < \hat{t}_{j+1} - tt_{x_j,x_{j+1}}$ and that $t_j^+ + tt_{x_j,x_{j+1}} < \hat{t}_{j+1}$, i.e., the end time t_j^+ and the start time t_{j+1}^- have values in intervals larger than zero. The end time t_j^+ is uniformly distributed, $t_j^+ \sim U(\hat{t}_j, \hat{t}_{j+1} - tt_{x_j,x_{j+1}})$, with density function

$$f(t_j^+) = \frac{1}{\hat{t}_{j+1} - tt_{x_j,x_{j+1}} - \hat{t}_j}.$$

The start time t_{j+1}^- is uniformly distributed between $t_j^+ + tt_{x_j,x_{j+1}}$ and $\hat{t}_{j+1}$. Since t_j^+ is itself uniformly distributed, the density function of t_{j+1}^- is

$$f(t_{j+1}^-) = \frac{1}{\hat{t}_{j+1} - tt_{x_j,x_{j+1}} - \hat{t}_j} \ln \frac{\hat{t}_{j+1} - tt_{x_j,x_{j+1}} - \hat{t}_j}{\hat{t}_{j+1} - t_{\psi+1}^-}$$

(and expected value is $E(t_{j+1}^-) = \frac{\hat{t}_j + tt_{x_j,x_{j+1}} + 3 \cdot \hat{t}_{j+1}}{4}$) (see Appendix A.1 for a derivation). In the cases when $\hat{t}_j \geqslant \hat{t}_{j+1} - tt_{x_j,x_{j+1}}$ or $t_j^+ + tt_{x_j,x_{j+1}} \geqslant \hat{t}_{j+1}$, t_j^+ and t_{j+1}^- are fixed with value $\hat{t}_j$ and $\hat{t}_{j+1}$ respectively.

3.3.3 Intermediary measurements

The duration of an activity episode is assumed to have a lower bound $T_{\min}$. Any episode with an expected duration $E(t^+) - E(t^-) < T_{\min}$ is rejected. It is assumed that the corresponding measurement has been generated while the pedestrian was walking, and therefore does not correspond to an activity episode.

Removing intermediary measurements also deals with outliers measurements. When a measurement error happens, a distant measurement is observed and an activity episode is generated. Because it is an outlier and it is not realistic, the distance to reach this activity episode from the previous one is long and therefore the travel time is long. The activity episode duration is consequently short and the wrongly generated activity episode is rejected.

In the case of very imprecise measurements, the DDR might be bounded for efficiency (see Section 3.3.2). It also avoids accumulated activity probability on a location with a strong prior based on no localization evidence (very weak measurement). It creates false activity episodes, artificially close to the measurement location. Then, if there is no confirmation from another measurement in this area, the time spent at this activity episode will be very short and thus this activity episode will be eliminated.

3.3.4 Sequence elimination procedure

The number of paths in the network grows exponentially with the number of measurements. For each measurement $\hat{m}_j$, $j = 1, ..., J$, all the elements of the corresponding DDR, $|DDR_j|$, have to be connected with all the previous candidates, resulting in $\prod_{j=1}^{J} |DDR_j|$ candidates. In practice, it is not possible to consider all possible combinations. Therefore, the proposed

implementation of the procedure imposes an upper bound L on the number of candidates. Whenever the number of candidates exceeds L, the least likely candidates (according to Equation 3.1) are eliminated to enforce the maximum number of paths in the network (Fig. 3.5). This procedure performs better deterministically (keeping the L most likely candidates) than stochastically (drawing L candidates based on the activity probability). Indeed, accumulation over several measurements generates an activity episode. When randomly picking a candidate, there is a risk to have several activity episodes with a small duration. They will be considered as intermediary measurements and eliminated.

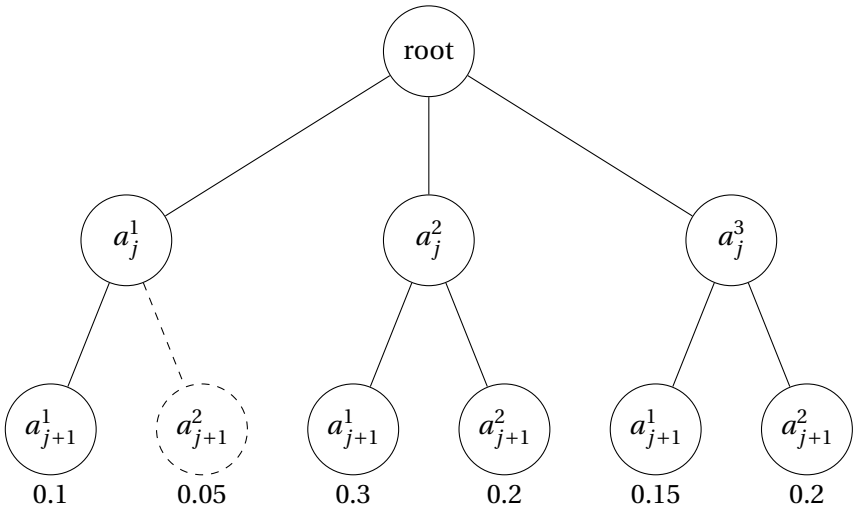


Figure 3.5 – Illustration of the sequence elimination procedure. The tree of Fig. 3.3 with the probability for each leaf to be the correct one. If $L = 5$, we eliminate the candidate represented by the dotted line, as it is associated with the lowest probability.

The processes described in Section 3.3.1, 3.3.2, 3.3.3 and 3.3.4 define Algorithm 1. It is illustrated in Fig. 3.6. The algorithm runs in $O(J \cdot |DDR| \cdot L \cdot |\mathscr{E}| \cdot |\mathscr{N}| \cdot \log(|\mathscr{N}|))$. Computational burden mainly comes from the shortest path algorithm. The number of shortest path computations depends on the size of the DDR (controlled by the modeler) and on the number of candidates L (also controlled by the modeler).

The network traces bring the dynamics in the process by allowing to track a pedestrian during all the journey in the pedestrian infrastructure. The prior is a way to add information about time constraints and attractivity. Finally, the pedestrian semantically-enriched routing graph (SERG) has two roles in the process. First, it allows to link the network traces (coordinates in a continuous space) to time constraints and attractivity of POI in the prior (places and landmarks in a discrete space). Second, shortest path in SERG being bigger than Euclidean distance between two POI, it corrects for anisotropy in the pedestrian infrastructure. It impacts the elimination procedure through the computation of shortest paths.

The proposed probabilistic measurement model computes the probability of performing an activity-episode sequence while generating measurements. It assumes that each measurement

Algorithm 1: Generation of activity-episodes sequences.

for *each ID* **do**
  for *measurement* $\hat{m}_j = (\hat{x}_j, \hat{t}_j)$, $j = 1, ..., J$ $O(J)$ **do**
    Define the corresponding Domain of Data Relevance, DDR_j $O(|DDR|\log(|\mathcal{N}|))$;
    for *each* $x \in DDR_j$ $O(|DDR|)$ **do**
      Compute the measurement likelihood ;
      if T *empty* **then**
        Initialize the network structure for activity-episodes sequences T with $x_1 = x$, $t_1^- = t_1^+ = \hat{t}_1$;
        Update the activity probability with the measurement likelihood and the prior ;
      else
        for *each* $a_{1:\psi}$ *of* T $O(L)$ **do**
          if $x_\psi = x$ **then**
            Update the definition of the episode end time: $t_\psi^+ = \hat{t}_j$;
            Update the prior for a_ψ ;
            Update the activity probability of $a_{1:\psi}$ with the measurement likelihood and the prior ;
          else
            Compute the shortest path between x_ψ and x, and the travel time $tt_{x_\psi,x}$ $O((|\mathcal{E}| + |\mathcal{N}|)log(|\mathcal{N}|))$;
            Define the last episode end time: $t_\psi^+ \sim U(\hat{t}_{j-1}, \hat{t}_j - tt_{x_\psi,x})$;
            Define the new episode start time: $t_{\psi+1}^- \sim U(\hat{t}_{j-1} + tt_{x_\psi,x}, \hat{t}_j)$;
            $a_{\psi+1} = (x, t_{\psi+1}^-, \hat{t}_j)$;
            if a_ψ *is an intermediary measurement* **then**
              Connect $a_{\psi-1}$ with $a_{\psi+1}$ in T ;
              Compute the prior for $a_{\psi+1}$;
              Update the activity probability of $a_{1:\psi+1}$ with the new measurement likelihood and prior, but without the prior for a_ψ ;
            else
              Connect a_ψ with $a_{\psi+1}$ in T ;
              Update the prior for a_ψ and compute it for $a_{\psi+1}$;
              Update the activity probability of $a_{1:\psi+1}$ with the new measurement likelihood and priors ;
    Sequence elimination procedure: keep the L most likely paths of the network T $O(L|DDR|\log(L|DDR|))$

corresponds to an activity episode. In reality, some measurements are generated while walking, and are then eliminated (see Section 3.3.3). If the measurements are dense enough, a possible extension of the proposed model consists in applying a probabilistic map matching approach such as Bierlaire et al. (2013) for measurements related to walking. Route choice models, and in particular for pedestrians, could then be applied and associated with the activity modeling.

3.4 A case study on EPFL campus

We conduct an experiment on the EPFL campus. We assume that the only mode on campus is walking, even if some people outside of the campus could be detected, either within a car on the road or within public transportation.

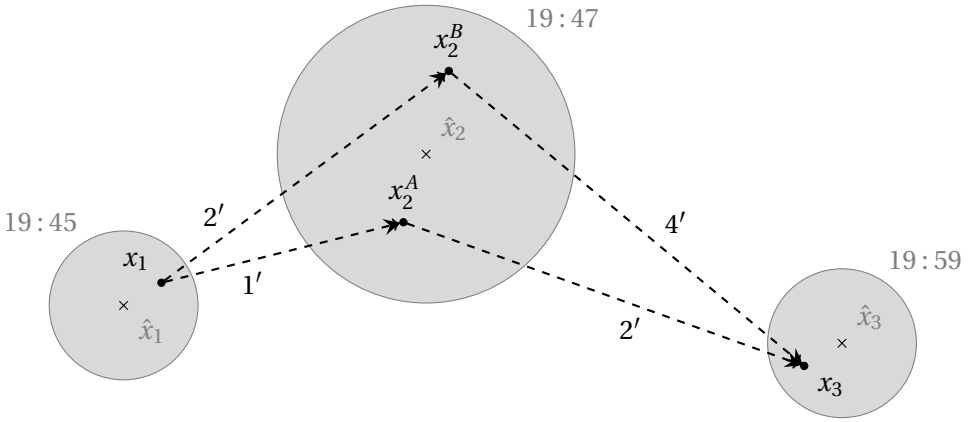


Figure 3.6 – Illustration of the full detection methodology, including the Bayesian probabilistic measurement model, the generation of activity episodes, intermediary measurements and sequence elimination procedure. The first measurement $\hat{m}_1 = (\hat{x}_1, \hat{t}_1)$ takes place at $\hat{t}_1$ = 19:45. Its DDR contains only one POI, x_1. It is connected with the two elements x_2^A and x_2^B of the DDR of measurement $\hat{m}_2 = (\hat{x}_2, \hat{t}_2 = 19{:}47)$. At 19:47, there are two candidates, (a_1, a_2^A) and (a_1, a_2^B). Let's assume a_2^A is twice more attractive than a_2^B, and x_2^A and x_2^B are in the same distance of the measurement location $\hat{x}_2$. Thus, (a_1, a_2^A) is twice more likely than (a_1, a_2^B). If $L = 1$, only (a_1, a_2^A) is kept, and then associated with x_3. With travel times from the picture, start and end times for a_2^A are generated: $t_2^- \sim U(19{:}45 + 1', 19{:}47)$, $t_2^+ \sim U(19{:}47, 19{:}59 - 2')$. Estimated time spent at a_2^A is 5'30. If $L = 2$, both (a_1, a_2^A) and (a_1, a_2^B) are kept and associated with x_3. Then, start and end times for a_2^B are generated: $t_2^- \sim U(19{:}45 + 2', 19{:}47)$, $t_2^+ \sim U(19{:}47, 19{:}59 - 4')$, for an expected time spent at a_2^B of 4'. As it is less than 5', a_2^B is eliminated and the two candidates are now (a_1, a_2^A, a_3) and (a_1, a_3). (a_1, a_3) is the most likely sequence since the measurement likelihoods are the same but the priors are $P(a_1, a_3) > P(a_1, a_2^A, a_3)$.

In Section 3.4.1 and 3.4.2, localization data and a pedestrian semantically-enriched routing graph of the campus are presented. We show how they comply with the data requirement defined in Section 3.2. Then, in Section 3.4.3, the potential attractivity measure used to generate the prior distribution is described. Results are presented in Section 3.4.4. Finally, in Section 3.4.5, sensitivity analysis is performed on the different parameters, in particular the ones defining the DDR, the prior and the density of measurements.

3.4.1 EPFL WiFi data

The data used for this case study have been collected with the Cisco Context Aware Mobility API with the Cisco Mobility Services Engine (MSE) (Cisco; 2011). Based on the signal strength from the 789 existing access points on campus, it uses triangulation to generate a measurement $\hat{m} = (\hat{x}, \hat{t})$. Therefore, in this case study, the location of the device $\hat{x}$ is continuous in space. A confidence factor cF defines a square around each x-y coordinates (see Fig. 3.7 and Cisco; 2011). The device is estimated to be inside this confidence square centered at the measurement $\hat{x}$ with sides $2 \cdot cF$ with 95 % probability. cF is calculated assuming that the device is located on the correct floor (Cisco; 2011). These data correspond to the localization data requirement

defined in Section 3.2.1.

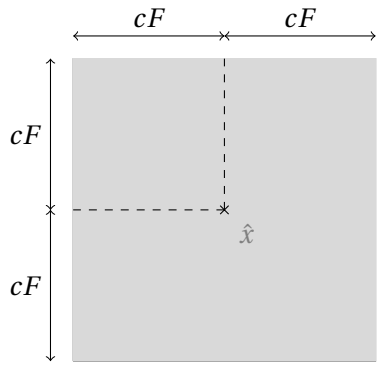


Figure 3.7 – 95 % confidence square provided by the localization tool (Cisco; 2011).

For the measurement equation as defined in Section 3.3.1, we assume that the errors in latitude and longitude are independently and normally distributed. We decompose both the measurement $\hat{x}$ and the activity location x in latitude and longitude $\hat{x}_{lat}, \hat{x}_{long}, x_{lat}$ and x_{long}. Assuming the errors in latitude and longitude are independent, $P(\hat{x}|x) = P(\hat{x}_{lat}|x_{lat}) \cdot P(\hat{x}_{long}|x_{long})$ with:

$$P(\hat{x}_{lat}|x_{lat}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\hat{x}_{lat} - x_{lat})^2}{2\sigma^2}\right) \tag{3.10}$$

$$P(\hat{x}_{long}|x_{long}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\hat{x}_{long} - x_{long})^2}{2\sigma^2}\right) \tag{3.11}$$

where $\sigma = \frac{cF}{2}$. This is equivalent to assume a Rayleigh distribution for the distance between the measurement $\hat{x}$ and the activity location x (Chen; 2013).

We collected 2'392'973 network traces (Fig. 3.8). The raw data are available in Danalet (2015). The confidence factor has a mean of 206 meters, with a minimum of 16 meters and a maximum of 1480 meters. The distribution of cF is shown in Fig. 3.9. Fig. 3.10 shows the spatial distribution of cF. Measurements with low precision are mostly outdoor.

3.4.2 EPFL pedestrian semantically-enriched graph

The EPFL website proposes an orientation tool for the campus, `http://map.epfl.ch`. It provides locations of offices and points of interest (such as restaurants and classrooms) on campus (Fig. 3.11). It also generates itineraries between two such locations. It consists of a semantically-enriched graph (SERG) as defined in Section 3.2.2, containing $|\mathcal{N}| = 50131$ nodes, $|\mathcal{E}| = 56655$ edges, and $|POI| = 5387$ points of interest.

The network as described above corresponds to the minimum data requirement as defined

Figure 3.8 – EPFL WiFi data (background: ©OpenStreetMap contributors, CC BY-SA).

in Section 3.2.2. However, more information is provided. Similarly to road networks for car driving, each edge is associated with a hierarchical status. Based on this hierarchical status, weights are defined in the routing tool of EPFL website as shown in Algorithm 2. The higher the weight is, the less likely the link is to be selected for the shortest path.

3.4.3 Potential attractivity measure on campus

On campus, each point of interest, $x \in POI$, belongs to one of seven categories: offices, classrooms, laboratories, restaurants, shops, library, and other points of interest. For each POI, we define attractivity $att(x,t)$ depending on the category it belongs to.

For each office, attractivity is equal to the aggregate work rates of employees provided by the human resources management software. For classrooms, attractivity equals the number of students who subscribed for a course at the beginning of the semester. This number varies with the time of the day. For restaurants and the library, we use the number of seats as a proxy. For shops on campus, no information is available and we arbitrarily assume that attractivity corresponds to a capacity of 20 people. Finally, for all other points of interest, we arbitrarily assume an attractivity of one, since we have no information about it.

Time constraints $\delta_{x,n}(t)$ as defined in Section 3.3.1 are based on class schedules for classrooms,

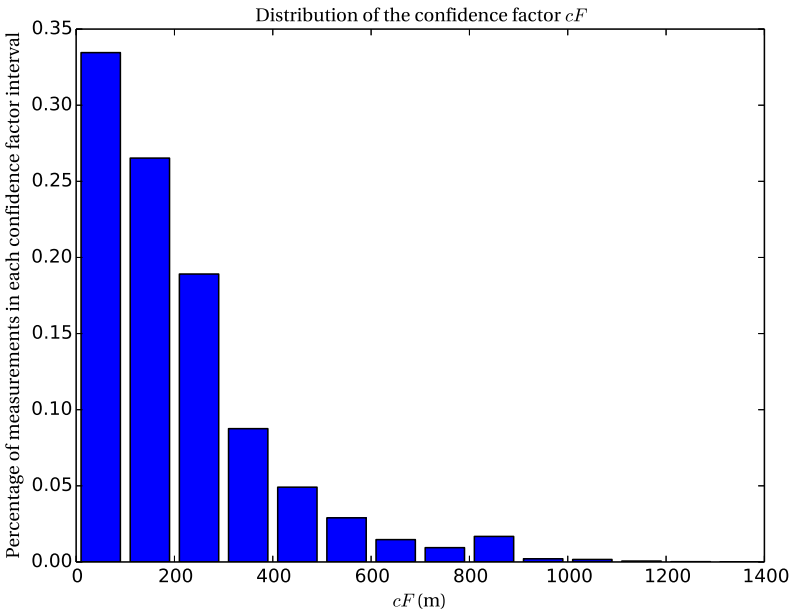


Figure 3.9 – Histogram of the different values of the confidence factor cF (in meters), based on 2'392'973 measurements.

and opening hours for restaurants, the library and shops. For offices, we assume no time constraint, and thus $\delta_{x,n}(t) = 1 \; \forall t$.

3.4.4 Results

Knowing the actual activity-episode sequence for the author

The methodology presented in Section 3.3 is tested with traces from the author. 76 measurements were generated on Monday May 14, 2012 (Fig. 3.12).

With EPFL WiFi data, a confidence square is defined assuming that the device is located on the correct floor. In order to account for floor error, we also consider the below and top floors, using a square with side $2 \cdot r$ on these floors. We define F as the probability of being in the detected floor, and $\frac{1-F}{2}$ the probability of being on the below or top floor. Both r and F are not provided and must be fixed by the modeler.

With this definition and given the high density of potential episode locations in the pedestrian network (in particular for offices, see Fig. 3.11), the number of locations in DDR is large (a mean of 712.0 with $r = 25$ m). The most distant episode locations of the DDR have a very low measurement likelihood. Decreasing the size of the DDR decreases the computation time. Moreover, very weak measurements generate huge DDRs and create a risk of giving too much importance to the prior. We define a maximum distance R in meters for taking potential

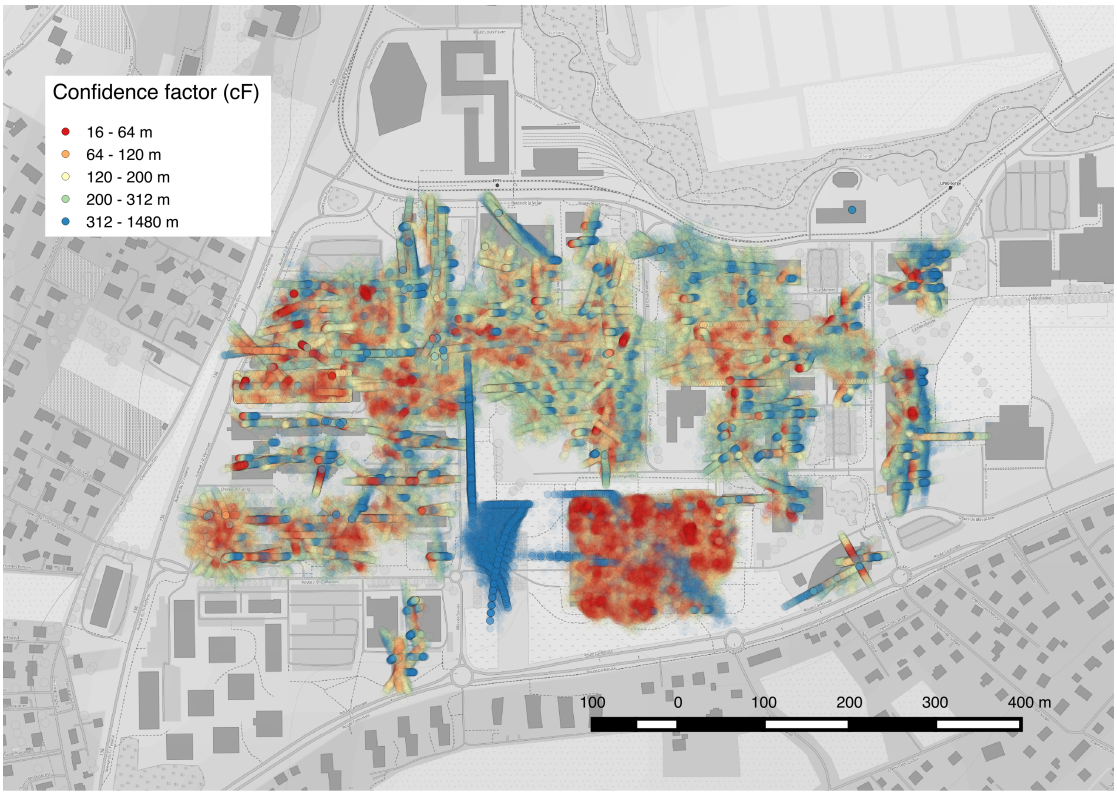


Figure 3.10 – Spatial distribution of the different values of the confidence factor cF (each point has a 95 % transparency; background: ©OpenStreetMap contributors, CC BY-SA).

episode locations in consideration, $cF_{\text{trunc}} = \min(cF, R)$.

The sequence of activity episodes has been recorded manually by the author and is shown in Table 3.1 and Fig. 3.13. He first went in a classroom from 8:32 to 10:30 for a course, then in his office until 11:47. For lunch break, he arrived in a restaurant on campus at 11:55. He came back to his office around 13:00 and went for a coffee around 14:00. Finally he came back in his office until the end of his working day, around 19:45.

The results are presented in Table 3.2, with $R = 80\,\text{m}$, $r = 25\,\text{m}$ and $F = 0.6$. At each iteration, only the best candidate is kept ($L = 1$). The potential attractivity measure is using the individual disaggregate class schedules. Δx is the shortest walking path between the episode location from the model and the one from the activity log in the semantically-enriched routing graph.

Compared to the mean confidence factor $\overline{cF} = 124.2\,\text{m}$, the spatial error ($\Delta x$) is low. The last activity episode, the metro stop, is not covered by WiFi. It is at the border of the campus. Thus, the error is big in this case. 3 out of 7 activity episodes are perfectly detected, and 3 more have a correct category. The number of episodes is correctly detected, as well as the floor of each activity episode. The temporal precision seems coherent with the diary. Results are presented on a map in Fig. 3.14.

Figure 3.11 – EPFL pedestrian semantically-enriched graph with points of interest and pedestrian network.

As the precision of the WiFi data is low and attractivity measure does not perfectly correct for this imprecision, the number of candidates L in the sequence elimination procedure (Section 3.3.4) can be increased to represent this uncertainty. L must be defined by the analyst to balance between algorithm speed and representation of uncertainty in the data. Figure 3.15 shows results with $L = 100$ candidates. Some activity episodes are present in each of the 100 candidates, expressing the absence of ambiguity at this time of the day (episodes 3 and 5, restaurant and cafeteria). In other cases, a strong ambiguity, both in horizontal error and activity-episode category, is present (episode 1, classroom). Measuring this uncertainty allows for corrections in further analysis.

Individual and aggregate results for campus members

The same methodology was applied to 3490 employees and 767 students of campus (with $L = 20$ and $F = 0.99$). Data were limited to 5 weekdays, from May 17 to May 23, 2012. Campus users authenticate themselves on the WiFi network through WPA (WiFi Protected Access) using a Radius server. Accounting is one of the process on the Radius server. It allows to associate a MAC address with a username (Koo et al.; 2003). The username was associated with employee or class attribute through LDAP (Lightweight Directory Access Protocol) requests. Then, both

Algorithm 2: Weight definition procedure for each edge in the pedestrian network

```
if door = closed then
    weight = ∞;
else
    if Major Route then
        hierarchical factor = 1;
    else if Inter-building Route then
        hierarchical factor = 1.2;
    else if Intra-building Route then
        hierarchical factor = 1.5;
    else if Access to Offices then
        hierarchical factor = 2.0;

    floor factor = 1;
    if Up then
        if Ramp then
            floor factor = 3;
        if Stairs then
            floor factor = 15;
    if Down then
        if Ramp then
            floor factor = 2;
        if Stairs then
            floor factor = 12;

    lift factor = 0;
    if Elevator then
        elevator factor = 40;

    weight = length · hierarchical factor · floor factor + elevator factor;
```

Activity log		
Time spent	Floor	Location
8.32am-10.30am	1	Classroom
Until 11.47am	3	Author's office
From 11.55 am	1	Restaurant
Around 1pm	3	Author's office
Around 2pm	2	Cafeteria
Until around 7.45pm	3	Author's office

Table 3.1 – Sequence of activity episodes as reported by the author. It contains 6 activity episodes.

the MAC address and the username were deleted to guarantee privacy. This process generates anonymized network traces with known category of users on campus.

Figures 3.16, 3.17 and 3.18 show the activity patterns of two employees and one student. The POIs are aggregated per category in these figures. Figure 3.16 shows an arrival on campus between 8:05 and 8:10. The employee visits two offices first, then a restaurant, then an office

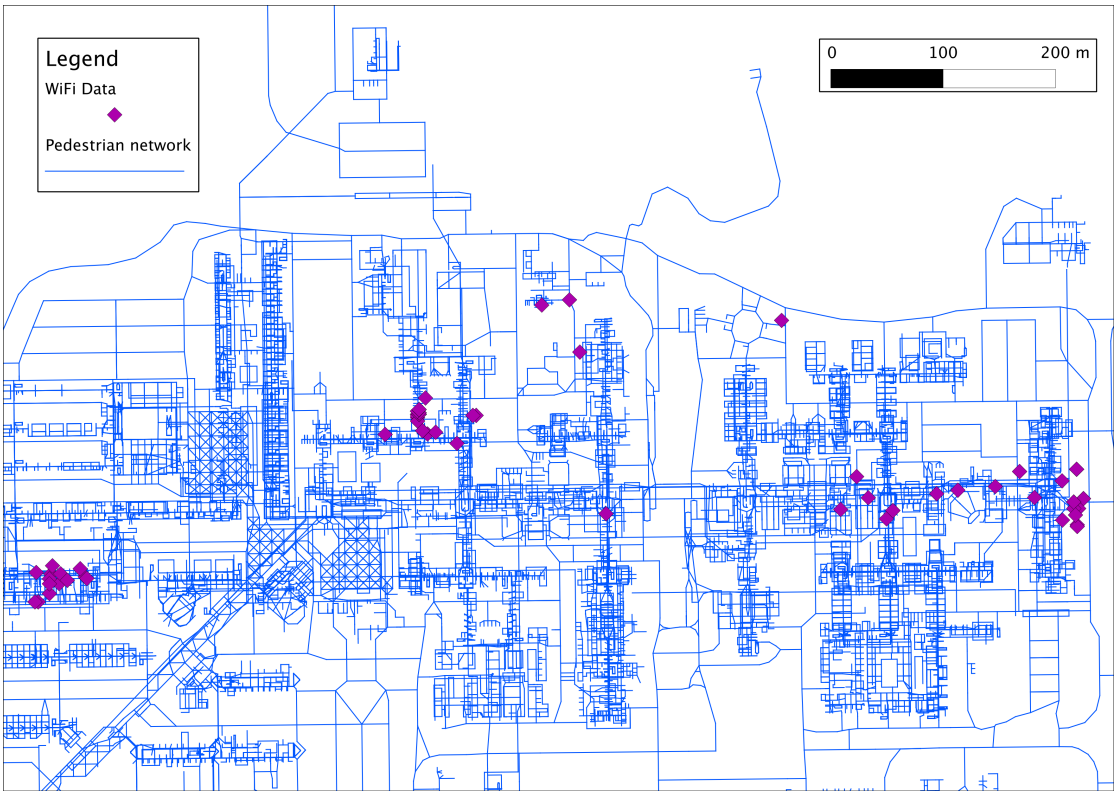


Figure 3.12 – WiFi traces generated by the author on Monday May 14, 2012 (violet) and the pedestrian network (blue).

Model with disaggregate prior				**Activity log**			Δx
Arrival time	Departure time	Floor	Location	Time spent	Floor	Location	(in m.)
8:35-8:35	10:38-10:38	1	Classroom	8.32am-10.30am	1	Classroom	0
10:40-10:40	11:51-11:51	3	Office	Until 11.47am	3	Author's office	9
11:54-11:54	12:47-12:53	1	Restaurant	From 11.55 am	1	Restaurant	0
12:51-12:58	13:03-13:44	3	Office	Around 1pm	3	Author's office	9
13:06-13:47	13:53-14:02	2	Cafeteria	Around 2pm	2	Cafeteria	0
13:55-14:04	19:40-19:44	3	Office	Until around 7.45pm	3	Author's office	9

Table 3.2 – Comparison between the most likely output of the model and the activity log as reported by the author.

and a lab in the morning. During lunch break, the employee visits two different restaurants. Then the employee visits a lab again, a restaurant, and finally the last episode is a lab with probability around 80 % and an office with probability around 20 %. Between 13:36 and 14:01, there is no destination where the measurements are stable for more than 5 min. This output seems realistic.

Figure 3.17 shows that WiFi devices are not necessarily mobile. Here, the device is accessing

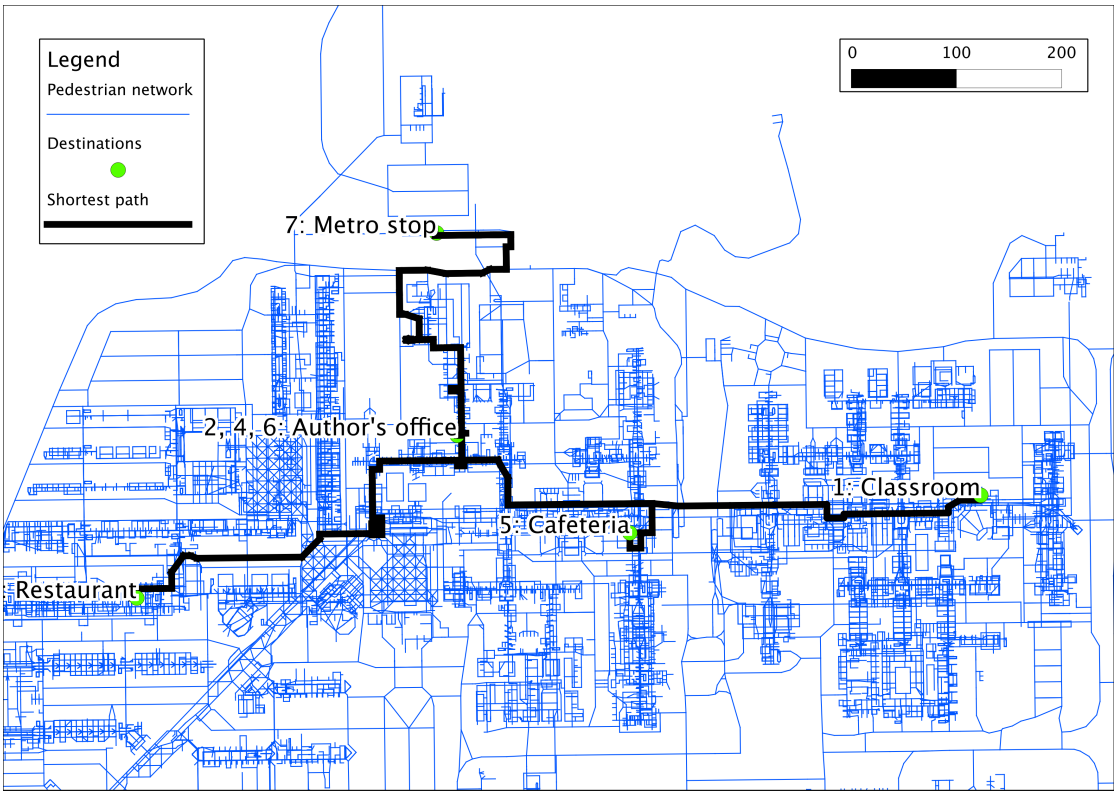


Figure 3.13 – Sequence of activity episodes as reported by the author. The dots represent activity episodes. The black thick lines represent the weighted shortest paths presented in Section 3.4.2. They use the pedestrian network and connect the activity episodes.

the WiFi all day long and not moving from one office. It is likely to be a fixed equipment. We remove such outputs in Appendix A.2.

A student's device activity pattern is shown on Fig. 3.18. The student's device was in a restaurant during lunch break and following courses in the afternoon. The "other" activity type in the morning represents here the campus bike service. It is very likely to be a measurement error since there is a class two floors up and 10 m away. Our measurement equation does not take into account more than one floor error, so the actual classroom is not in the domain of data relevance. The limitation of the size of the domain of data relevance increases the speed of the algorithm but also excludes some points of interest that could be realistic.

At a more aggregate level, we can observe from the output of Algorithm 1 that people on campus are performing 3 activity episodes on average. At an average they spend 1 h and 37 min on each activity. Focusing on the restaurants, Fig. 3.19 shows the number of devices detected in restaurants per quarter of an hour during the 5 weekdays. We observe a peak of transactions around noon, which is expected.

There are no data about the real behavior of people for validating the number of episodes, their duration or the proportion of people going to restaurants.

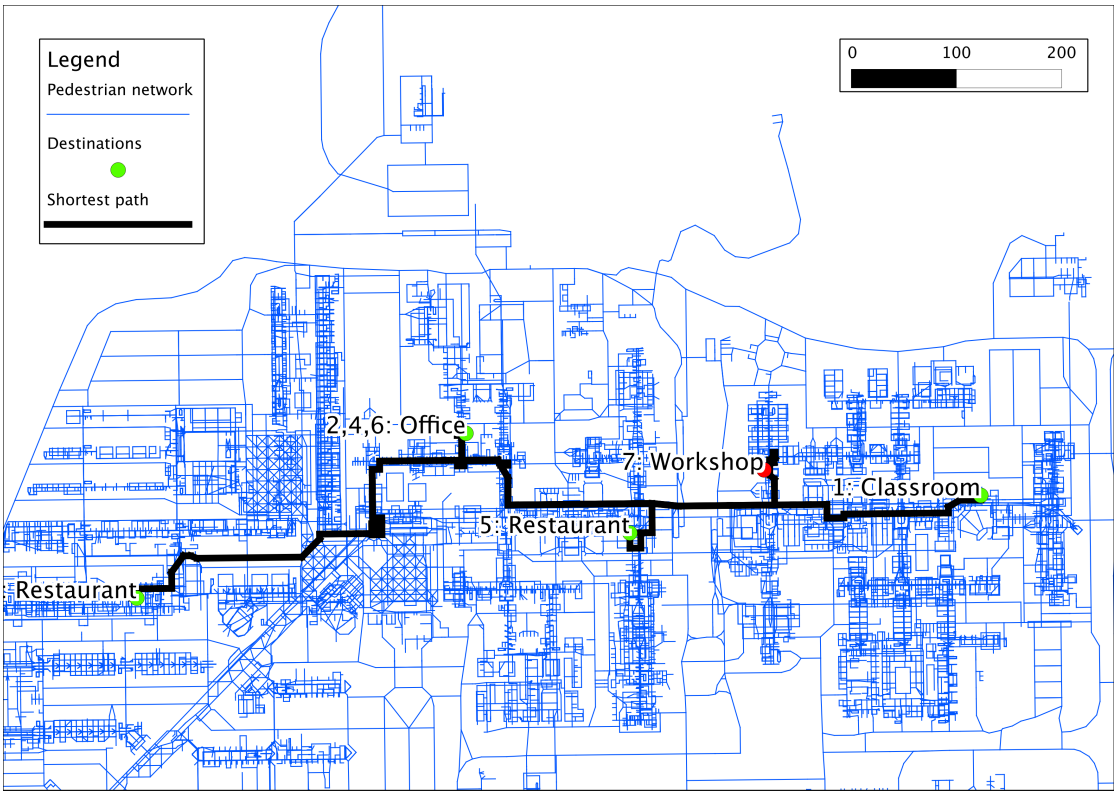


Figure 3.14 – Activity-episode sequence of the most likely output of the model with disaggregate prior on EPFL Campus pedestrian network. Episode locations are connected with the weighted shortest path presented in Section 3.4.2. Destinations are represented in green if in the correct category of POI and red otherwise. Only the last destination, the metro stop, not covered by WiFi, does not have the correct category.

3.4.5 Sensitivity analysis

Based on results from the author (Section 3.4.4), the sensitivity of the results to the parameters, the prior and the density of measurements are measured in terms of spatial and temporal precisions of each activity episode and more globally at the sequence level, quantitatively and qualitatively. Four criteria of stability are defined to evaluate the impact of the changes: the number of episodes that are detected by the algorithm ("Nb episodes"), the walking distance between the episode location from the model and the activity log one ("Delta dist.", in meters), the mean absolute difference between the activity log schedule and the schedule defined by the model ("Delta duration", in minutes), and the number of correct destinations categories ("Nb OK"). The reported walking distance, "Delta dist.", is the shortest path between the episode location from the model and the activity log one in the semantically-enriched routing graph. This criteria is more relevant than Euclidean distance since a small difference in localization may have a big impact on the actual distance in the pedestrian graph for the tracked individual. The reported episode start and end times are not very precise, in particular in the afternoon, and thus we only consider here the 5 start and end times with a sufficient

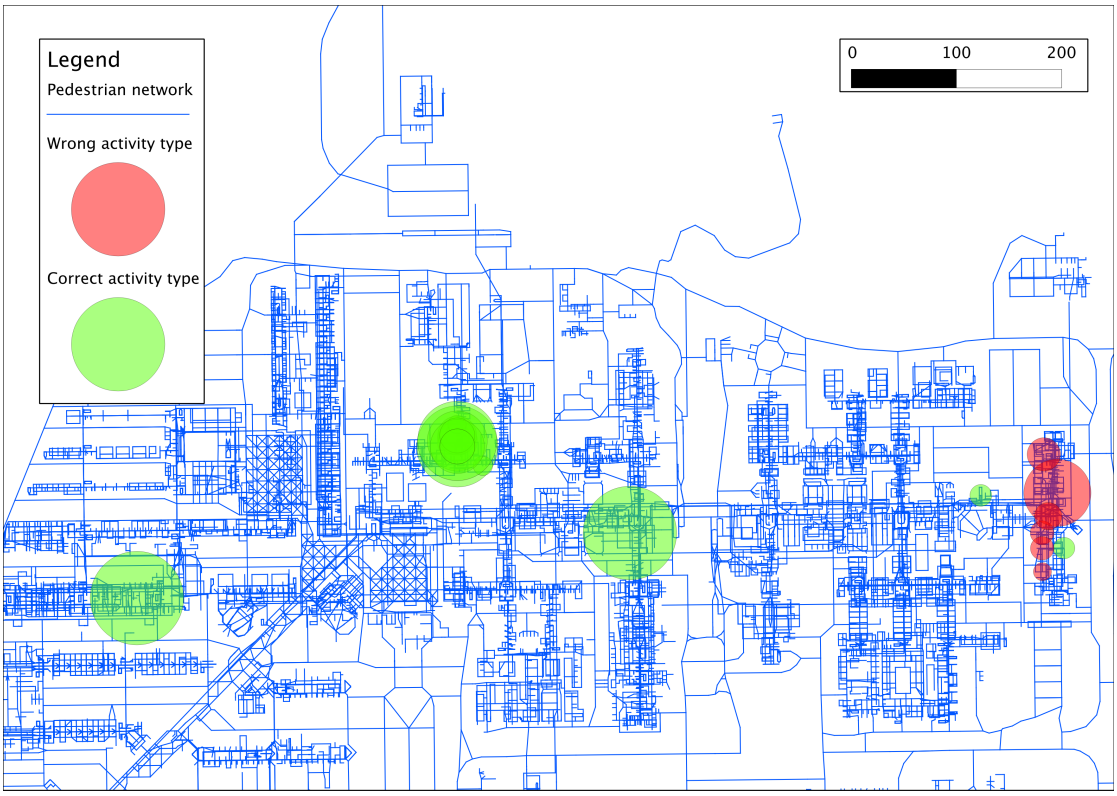


Figure 3.15 – Activity-episode sequence of the $L = 100$ most likely output of the model with disaggregate prior on EPFL Campus pedestrian network. Destinations are represented in green if in the correct category of POI and red otherwise. The surface of each point represents the normalized probability of this destination being the correct one (Equation 3.1). The two restaurants are detected in all 100 activity-episode sequences. The author's office is not always perfectly detected and variations can be observed, but the category is always correct. The classroom, in the beginning of the day, is not correctly detected, and the destination category is wrong in some cases. The actual classroom is detected in a minority of cases. In some of the 100 sequences, there is a seventh episode, but their likelihood is too small to be seen on the picture.

precision in the activity log schedule: $t_1^-, t_1^+, t_2^+, t_3^-, t_6^+$. The number of correct categories is important, since the detection of the exact office is not necessary for understanding mobility patterns, while knowing the kind of destination is crucial.

In the next sections, the impact of the changes in different parameters, the impact of the prior and the impact of the density of measurements are shown.

Sensitivity to the parameters

There are mainly four parameters that need to be defined in the model. First, the maximum radius R of the DDR, allowing to limit the computational burden related to some very imprecise

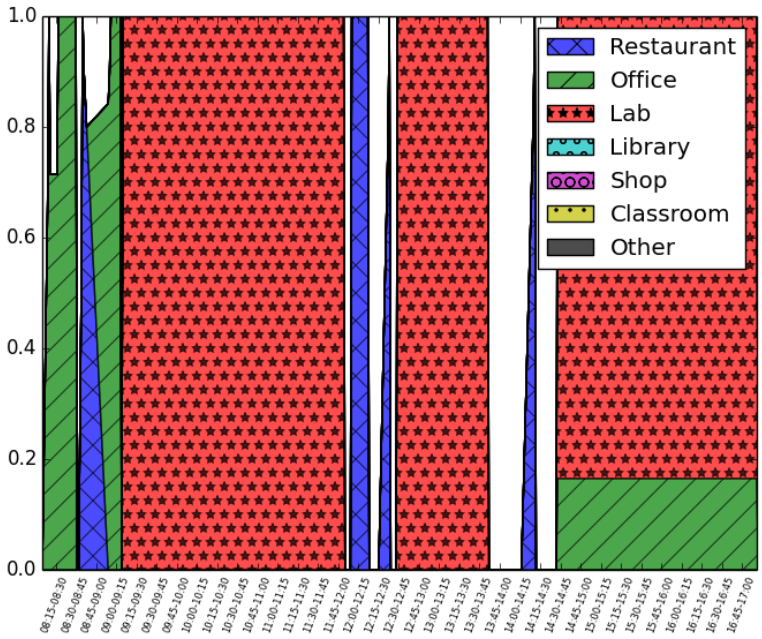


Figure 3.16 – Activity pattern for one employee's device on May 23, 2012. The x-axis represents the time of the day. The colors/patterns represent the different categories of the points of interest. The y-axis is the probability to be the correct point of interest based on Equation 3.1.

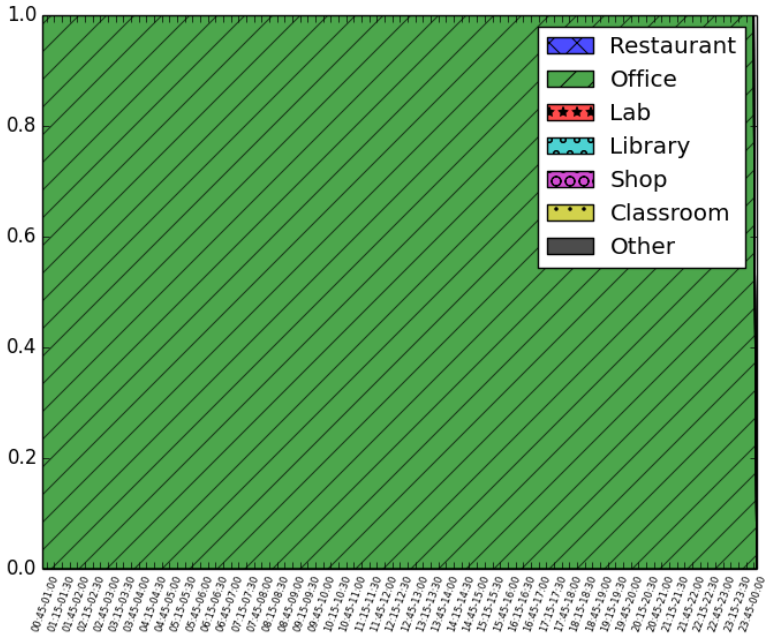


Figure 3.17 – Same representation than in Fig. 3.16 for another employee on the same day.

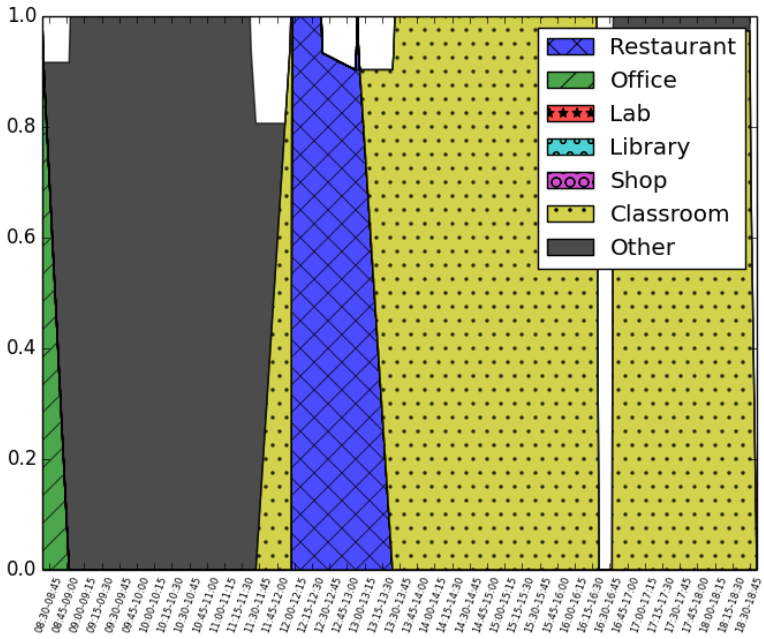


Figure 3.18 – Activity pattern for one computer science master student's device on May 23, 2012. The x-axis represents the time of the day. The colors/patterns represent the different categories of the points of interest. The y-axis is the probability to be the correct point of interest based on Eq. 3.1.

measurements; second, the probability F of being in the detected floor. Since the precision is expressed in the horizontal plan, a vertical precision needs to be set up; third, we also define the minimum time spent at destination T_{min} in minutes; and finally the number L of candidates that are kept during the sequence elimination procedure.

We use as a base case $R = 80\,\text{m}$, $F = 1.0$ and $T_{\text{min}} = 5\,\text{min}$, as in the previous results, and we fix $L = 40$. The expected values for the criteria of stability are used based on the normalized activity probability from Eq. 3.1.

Figure 3.6 showed the impact of L on a small illustration, with $L = 1$ and 2. With real data, the same effect is appearing for the last episode. It shows than $L = 1$ should be avoided (Fig. 3.20). By fixing the number L of candidates to 1 at each iteration, only the most likely last activity episode is kept at each measurement. This does not allow for explicit management of ambiguity of the measurement and does not provide memory to the process. With $L > 1$, results are stable.

Varying T_{min} defines the time length of intermediary measurements. For some large values of T_{min} $(9, 10, 11, 12\,\text{min})$, it is possible that performing several activity episodes of less than T_{min} is more likely than performing the actual activity episode. It explains the low values for "Nb episodes" and "Nb OK" in Fig. 3.21. Reasonable values for T_{min} represent the expected error

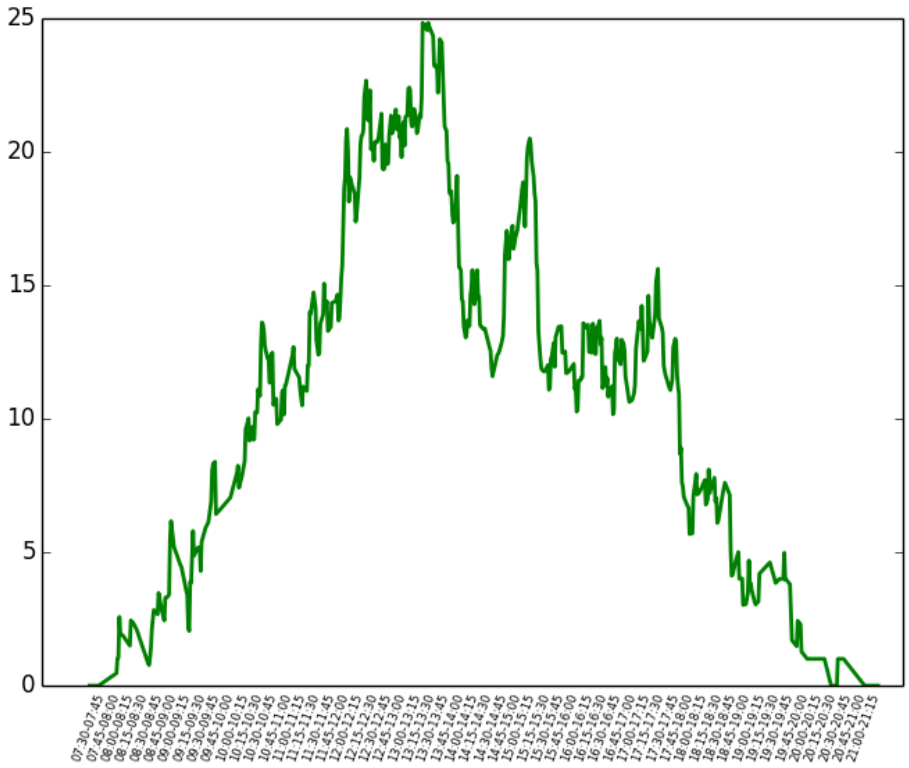


Figure 3.19 – Number of devices detected in restaurants per quarter of an hour. The x-axis represents the time of the day (quarters of hour). The y-axis represents the cumulated number of people detected in the restaurants of the campus over 5 days from the WiFi traces (devices/quarter). Since 20 activity-episode sequences are generated per individual, each one is weighted by its probability to be the correct one based on Eq. 3.1.

in the travel time between two episodes, because of slower walking speed or longer distance than the shortest path (see Fig. 3.22), i.e., less than 9 min. In these cases, results are stable. $T_{\min}$ should also be bigger than 1. In this case, very short and unrealistic activity episodes on the way between two actual activity episodes are more likely than the actual ones.

In our particular example, if R is small ($R = 30, 40$), the ambiguity for the first episode (see Fig. 3.15) disappeared. It is case specific. In general, with small R, destinations are missed by the algorithm. For large values ($R = 90, 100$) or no limitation ($R = \infty$), the geographical information provided by the WiFi measurement is almost flat. In this case, the activity probability depends on prior only and the prior is bigger for less activity episodes. It explains the low number of activity episodes in Fig. 3.23. The output is stable for $R = 60, 70$ and 80 m.

In our example, 10 of the 76 measurement are not on the correct floor (13 %). Only one of them corresponds to an activity episode (the 9 others are measurements related to the metro stop, not covered). Figure 3.24 shows that the interfloor probability F has a very small impact

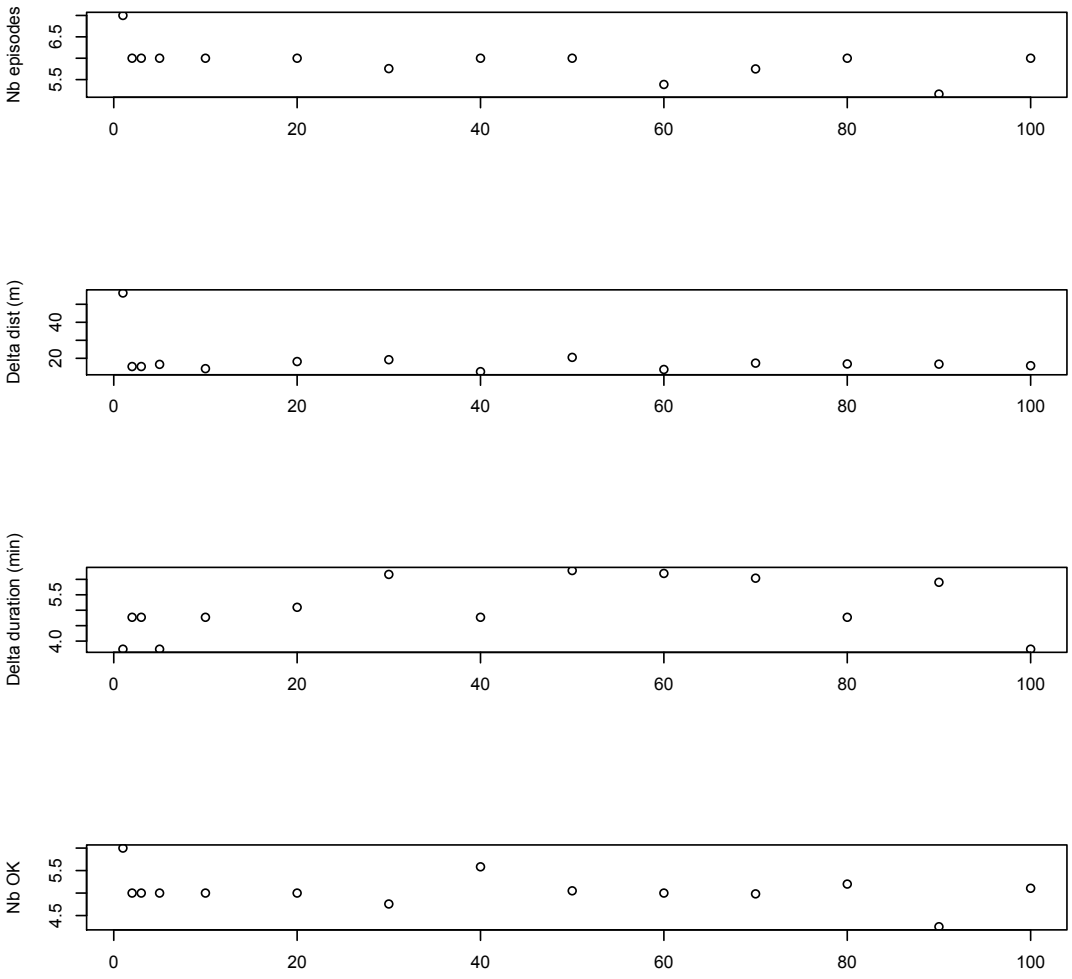


Figure 3.20 – Sensitivity to the number L of candidates kept between each measurement, $L = 1, 2, 3, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100$.

in this specific example, if F is in the order of magnitude of the error (0.9). Still, the vertical imprecision in a multifloor environment must be taken into account. In particular, when the device is next to windows, stairwells or mezzanines, the signal could cross the floor separation.

As an extra example, the author had a class on March 27, 2012. 2 of the 14 measurements corresponding to this activity episode are detected on the floor below the actual episode location (Fig. 3.25). These two consecutive measurements happened in more than 5 min difference and thus are not considered as intermediary measurements. Moreover, they happen after the beginning of the activity episode. With $F = 1$, three activity episodes are detected: on the correct floor, then downstairs, then on the correct floor again. With $F = 0.9$, only one activity episode is detected.

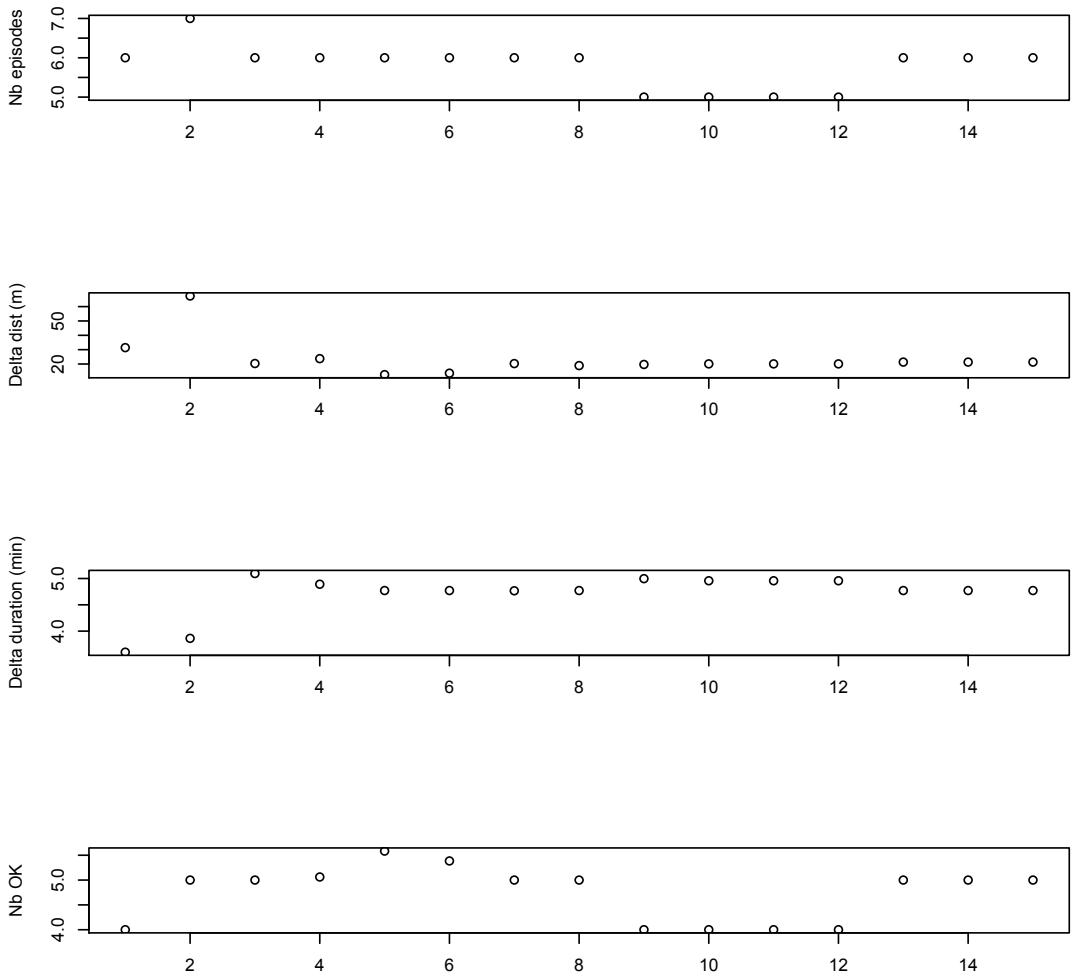


Figure 3.21 – Sensitivity to the minimum time spent at destination T_{min}, $T_{\text{min}} = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15$ min.

Effect of the prior

In Fig. 3.26, we present results with the different priors defined in Section 3.3.1 to show their effects:

- uniform;
- aggregate for all campus members, using the same class attractivity for all students;
- disaggregate for a class, meaning that we know the exact class schedule for the tracked pedestrian; and

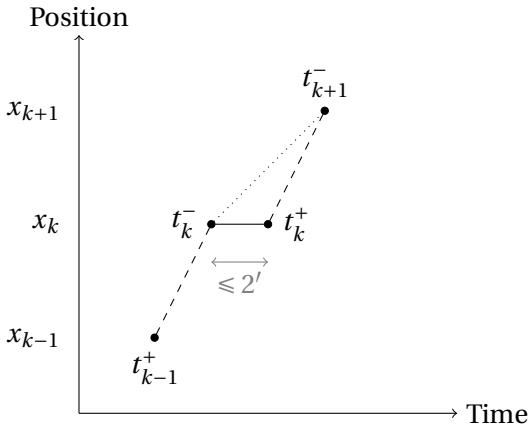


Figure 3.22 – Time-space representation of one activity episode ψ with short time spent at it, $t^+_\psi - t^-_\psi \leqslant 2$ minutes. The dashed line represents the assumed trip, with mean speed and shortest path. The dotted line is the actual trip, with slower walking speed or longer distance than the shortest path.

- diary, based on the recorded sequence presented in Table 3.1.

We observe first that the total number of episodes in the day is estimated correctly from the 76 measurements with each different prior (Fig. 3.26). Using a threshold of 5 min spent at episode locations, we reach the same number of episodes as the activity log. It means that without extra information, with a uniform prior, the WiFi data are already providing information about the number of episodes in the day. On the other hand, with a uniform prior, only 4 out of 7 activity episodes have correct category. This information is crucial for understanding and modeling activity choice.

We can observe that the aggregate prior is not precise enough (Fig. 3.26). The number of episodes is stable, as well as the number of correct categories ("Nb OK") compared to a uniform prior, and spatial precision is worse. The aggregate prior does not improve the results. A deeper analysis of the results shows that including class schedules for all pedestrians, even those to whom these schedules are not relevant, is giving too much importance to classrooms compared to offices and other points of interest. It creates a bias towards classes by applying the same time constraints to everyone, even when these schedules are wrong for a particular individual.

Applying the correct time constraints needs class schedules, and thus lower anonymity level of the WiFi data. The disaggregate prior does not require the student identity but to which class the student belongs to. It detects almost all destinations perfectly, with correct categories. The individual anonymity is kept, while the attractivity and time constraints allow to correctly detect the category of the episode. Spatial error ("Delta dist.") is almost as good as the diary prior.

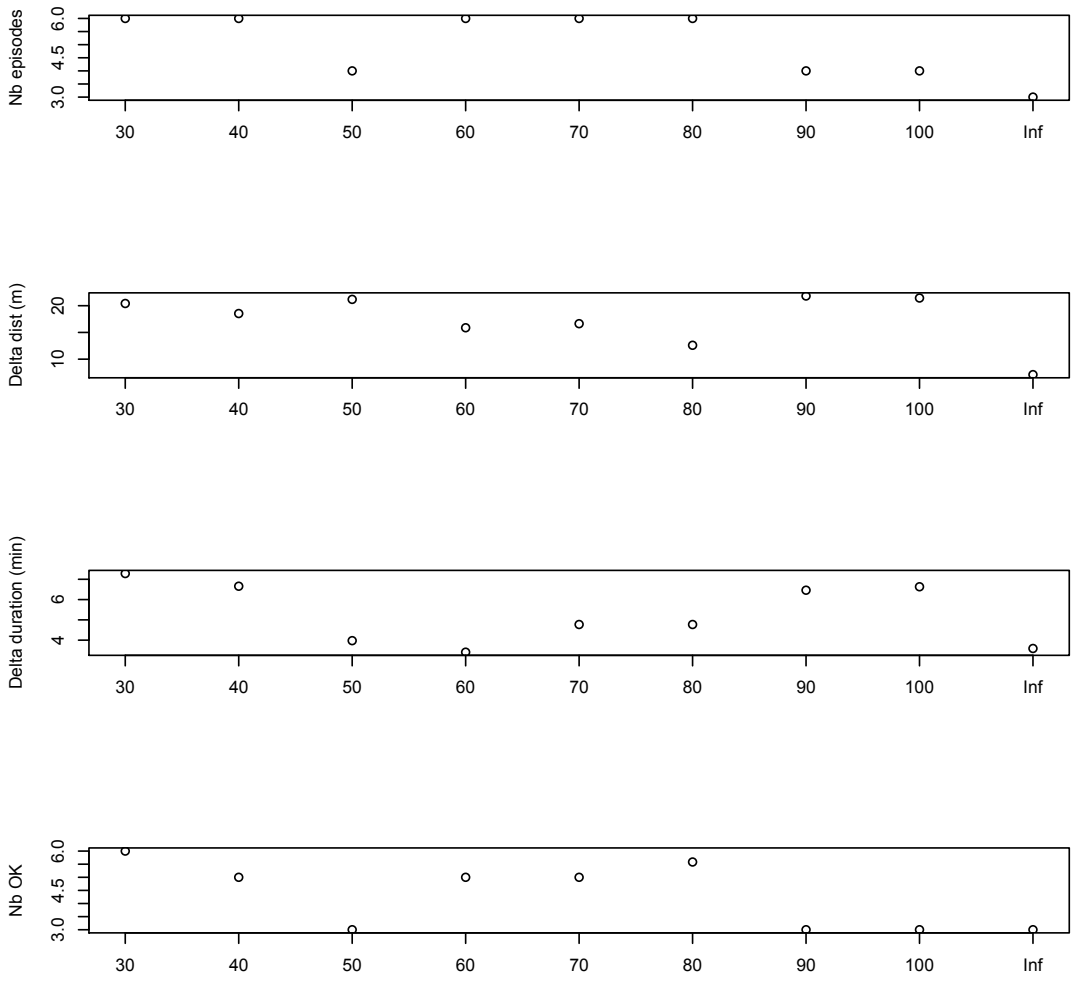


Figure 3.23 – Sensitivity to the maximum radius R of the DDR, $R = 30, 40, 50, 60, 70, 80, 90, 100$ meters and $R = \infty$.

The diary prior allows to correctly detect only 6 out of the 7 activity episodes since the metro stop is not covered by WiFi and out of the confidence square. It corresponds to the best possible results with our approach, when the value of the prior is 1 for the point of interest corresponding to the correct activity episode location (i.e., the POI that was actually visited by the device of the tracked individual) and 0 for the other points of interest in the DDR (the POIs that were not visited by the device of the tracked individual). In practice, it is impossible to collect enough data in order to build a diary prior. The exact same results, i.e. the best possible results, are also reached with an attractivity of 3 for the point of interested corresponding to the correct activity episode location and 1 for the other POI. It shows the needed order of magnitude of the prior to overcome WiFi data imprecision in a pedestrian infrastructure with

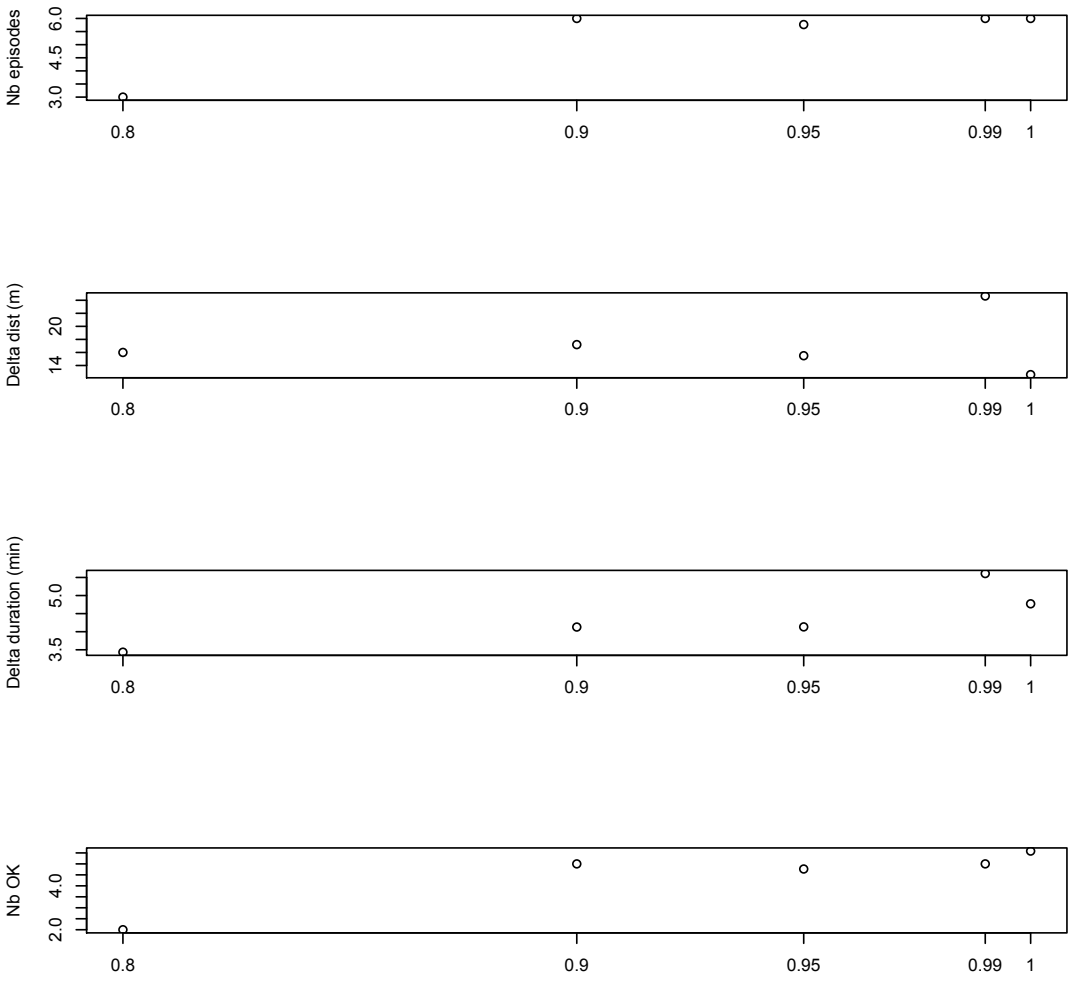


Figure 3.24 – Sensitivity to the probability F of being in the detected floor, and not in the upper or lower floor, $F = 0.8, 0.9, 0.95, 0.99, 1.0$. F=1.0 means that only the detected floor is considered.

dense points of interest and detect correct categories: the visited points of interest should have an attractivity 3 times larger than the attractivity of the non-visited POIs.

A prior with more information does not necessarily improve the results. Individual attractivity and time constraints allow to detect the correct categories of activity episodes and to reach a better spatial precision, while maintaining anonymity of the tracked pedestrians.

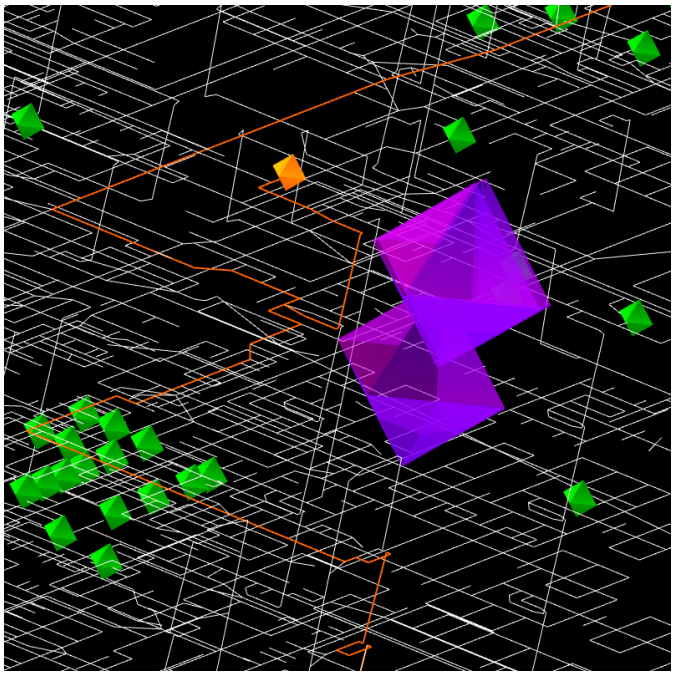


Figure 3.25 – Example of vertical imprecision. The violet diamonds represent the signals. There are 12 signals upstairs and 2 signal downstairs. The orange diamond is the location of the activity episode, on the upper floor. The orange line is the shortest path connecting the activity episodes. The green diamonds are classrooms. The white lines represent the pedestrian network. Offices are not shown for clarity (image: Lopez-Montenegro Ramil et al. (2013)).

Sensitivity to the density of signals

To evaluate the sensitivity of the model to the density of measurements, some data were artificially removed. 5 %, 10 %, 15 %, 20 %, 25 % and 30 % of the data were randomly removed. The model ran 100 times for each case. Results are shown in Fig. 3.27. We observe that the model is very stable until -15 % and then start to have variations, particularly in the difference of walking distance between the activity log and the model. This variation more frequently improves the difference of distance, which is good since the precision is better, but is less stable. Results with -25 % and -30 % of the full dataset show less stability, with variations in distances between the activity log and the models and also in the number of episodes and the number of correct destination categories.

As a general recommendation, 76 measurements cover properly an almost 12-h journey on a campus. Results are still stable and trustworthy with -15 % of measurements, i.e. 65 measurements, which corresponds to a mean of 5.4 measurements per hour.

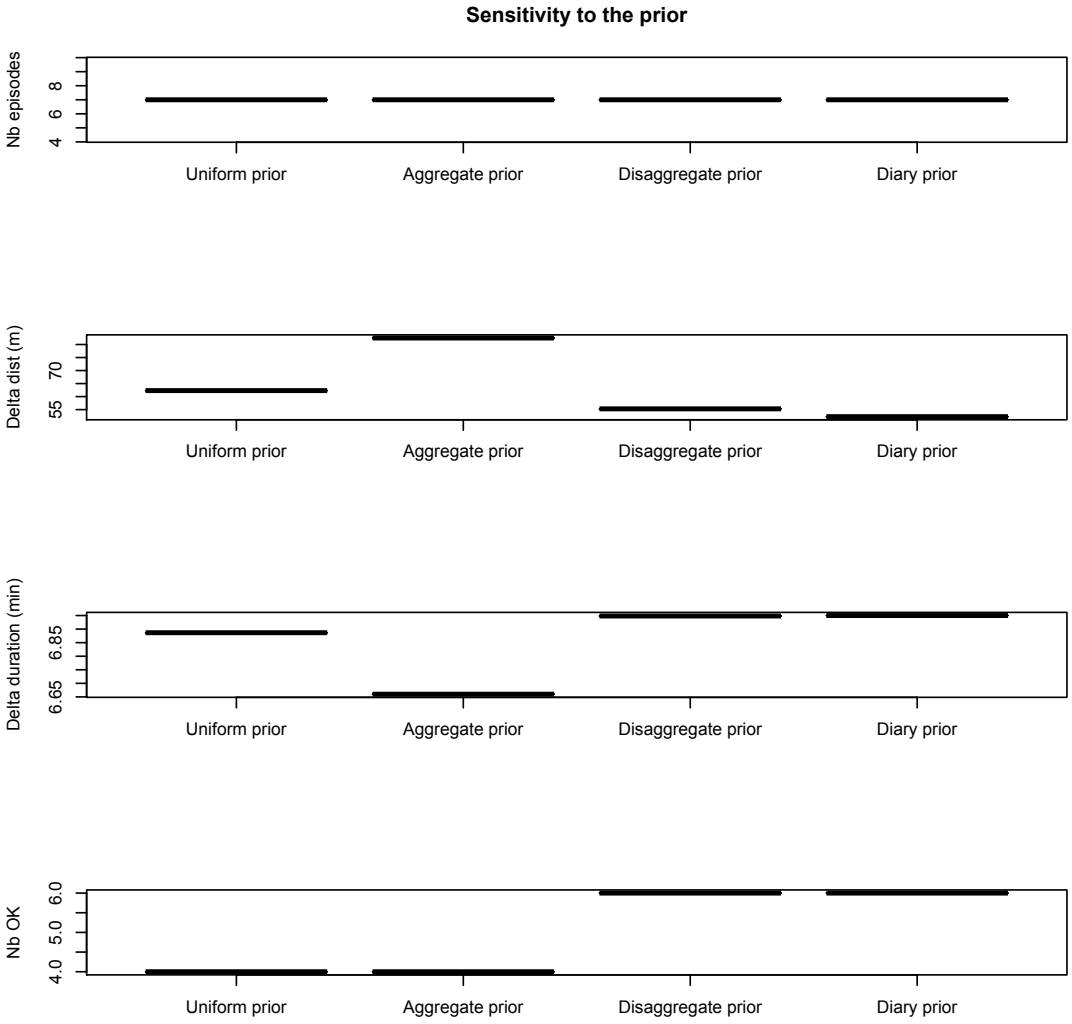


Figure 3.26 – Sensitivity to the prior, with uniform, aggregate, disaggregate and diary prior.

3.5 Conclusion

In this chapter, we propose a methodology detecting the different activity locations visited by a device using its network traces supported by knowledge of the underlying pedestrian map and attractivity, in particular time constraints. Semantics is extracted from raw data: activity-episode sequence, with start and end times, activity type and destination. We present an empirical study on a campus.

Our approach accounts for the fact that pedestrian networks are traditionally denser than other mobility networks and localization is often sparse, in particular indoor. The methodology presented here is flexible and tunable. It allows for introducing *a priori* knowledge on the activities and information on the pedestrian map structure. In particular, time constraints

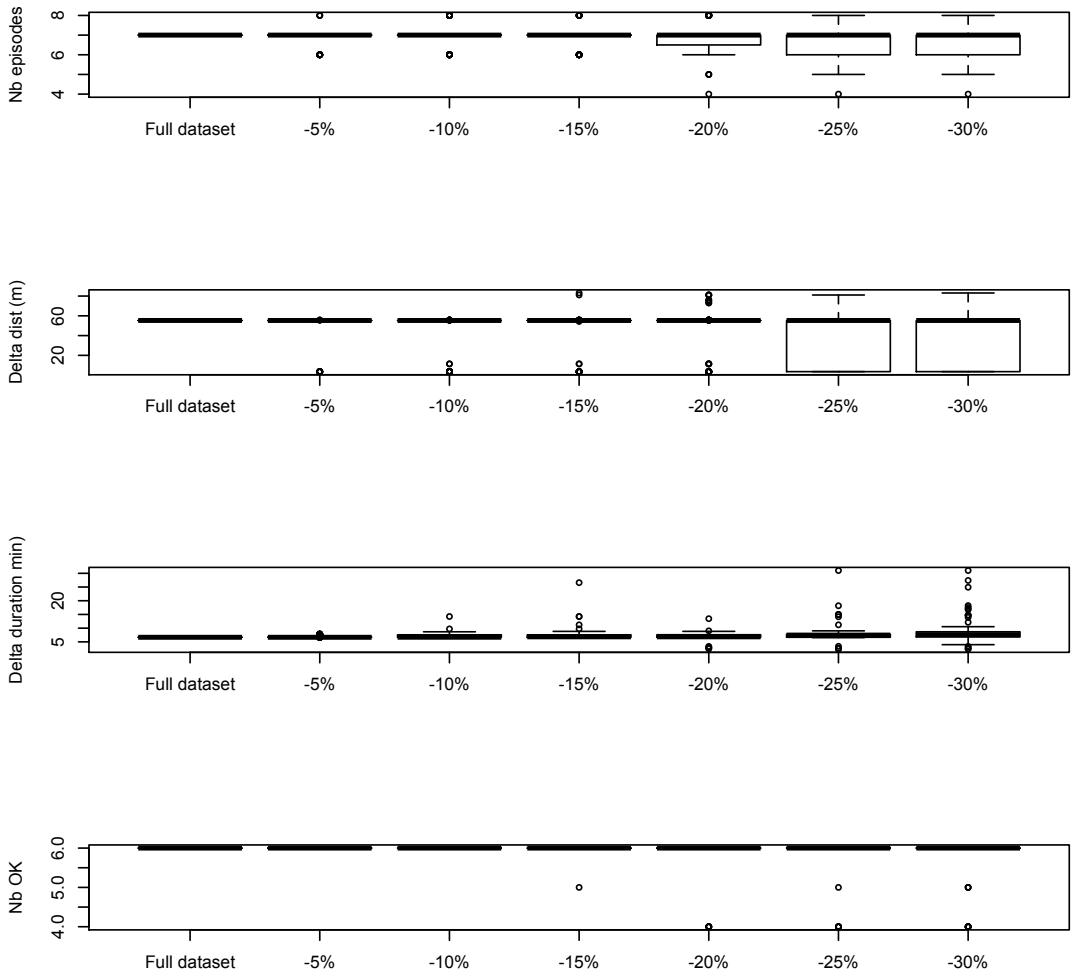


Figure 3.27 – Sensitivity to datasets with less data, with the full dataset as a base case.

(such as schedules for trains in a railway station, for planes in an airport, or for classes on a campus, or opening hours for shops or restaurants) can be added in the model. Moreover, the usage of a pedestrian network corrects for anisotropy in the facility.

This methodology also uses the concept of domain of data relevance. By using domains of data relevance, if access points are changing very often from one to another while the device is in fact static, the true activity location is contained in both domains of data relevance and does not change. Therefore, this methodology avoids the pingpong effect observed in other studies. The avoidance of the pingpong effect is reinforced by the prior, focusing on specific points of interest in the domain of data relevance.

This methodology is robust for low density measurements. Finally, ambiguity is explicitly

stated through the likelihood of each activity-episode sequence.

This approach has limitations. First, it works in pedestrian facilities and does not account for mode detection. Also, we emphasize the importance of a good knowledge of the map behind the technical infrastructure. As results show, more data in the prior does not necessarily mean better results, and a careful definition of attractivity and time constraints is needed. Finally, R, the bound for the size of the DDR, must be fixed by the analyst and may cause a wrong number of detected episodes.

The representation of points of interest may have an impact on the detection of activity episodes. When they are represented as points and for points of interest with a large surface, the measurement may take place in the facility but its domain of data relevance does not intercept the point representing the point of interest. In the case study on campus, offices, labs and classrooms are represented as surfaces. Restaurants, shops, libraries and other points of interest are represented as points. Figure 3.28 exemplifies the problem in EPFL library in the Rolex Learning Center.

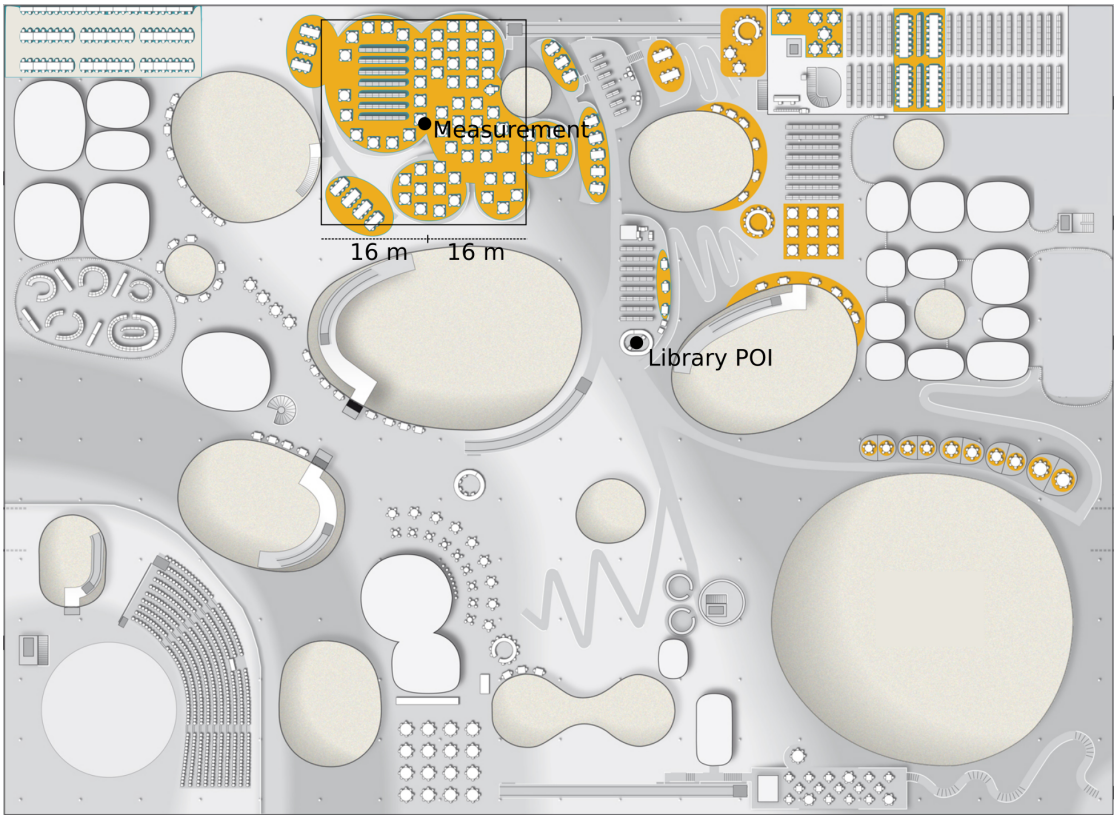


Figure 3.28 – Map of the Rolex Learning Center, with the library in the top right corner of the building. The point of interest “Library” is represented as a point, but the surface of the library is large (working tables in yellow) and it is possible to be detected in the library (measurement with $cF = 16\,\text{m}$) and not be associated with the “Library” point of interest (background: EPFL Library image)

Future works involve applying this methodology with different sensors and in different contexts. It can be used with other network traces, such as Bluetooth tracking. In other contexts, such as train stations, hospitals, festivals or airports, attractivity measures and time constraints are different. Also, more data can be included in the model, regarding the measurement equation, the prior, the shortest path, the first and last activity episodes, and the estimation of the parameters. The measurement equation may be improved by determining the source of propagation errors such as obstacles or walls. The prior could be further extended with models on activity choice or with more precise data about attractivity. The shortest path algorithm may describe big obstacles to increase length-optimality of the shortest path algorithm, or consider one-way paths similarly to street networks. The first and last episodes in the studied area are particular in the sense that they represent the access to the area. In our experiment, access to campus can be detected using prior knowledge, like studies about mode choice to access campus. The methodology can be extended by using the probabilistic map matching method developed by Bierlaire et al. (2013) for intermediary measurements. If exact activity-episode sequences are available for a large sample, e.g., from surveying pedestrians or from cameras, they can be used for Bayesian estimation of parameters such as F and r.

Modeling Part II

4 A path choice approach to activity modeling

4.1 A general model for activity-episode sequences

Our approach to activity-based travel demand modeling decomposes the modeling of behavior in two steps (see Fig. 1.1, page 5). First, a path choice approach models how people choose their activity type in time, taking into account the type of activities, their sequence and their timing/duration. Once the activity type, sequence and timing are chosen, a second step consists in modeling destination choice (e.g., choosing a restaurant, knowing the activity type: eating).

There are mathematical and behavioral motivations for this decomposition of behavior modeling in activity and destination choices. Mathematically, the problem is complex. The number of possible destinations in a given area is usually large. The number of sequences of destinations is larger. Including the duration spent at destination makes the problem definitively too large and intractable. Behaviorally, the choice of activity type and time of day precedes the choice of destination (e.g., Bowman and Ben-Akiva; 2001; Arentze and Timmermans; 2004; Abou-Zeid and Ben-Akiva; 2012; Kang and Recker; 2013).

Formally, we model the sequence $a_{1:\Psi}$ of activity episodes $a_1, ..., a_\psi, ..., a_\Psi$, where an activity episode $a_\psi = (x_\psi, t^-_\psi, t^+_\psi)$ is defined as a location x_ψ, a start time t^-_ψ and an end time t^+_ψ. The probability of reproducing the observation of a sequence of J measurements $\hat{m}_{1:J} = \hat{m}_1, ..., \hat{m}_j, ..., \hat{m}_J$ of individual n is decomposed as a measurement equation $P(\hat{m}_{1:J}|a_{1:\Psi})$ and the probability to choose an activity-episode sequence, $P(a_{1:\Psi})$ (Eq. 4.1). The measurement equation computes the probability that the performed episodes generated the observed measurement sequence. By assuming localization error only, the measurement equation can be decomposed as a product of the localization error for each measurement location $\hat{x}_j$, $\prod_{j=1}^{J} P(\hat{x}_j|x^j_\psi)$, where x^j_ψ is the activity episode location corresponding to measurement $\hat{m}_j$. The probability $P(a_{1:\Psi})$ of performing an activity-episode sequence $a_{1:\Psi}$ is decomposed in a model $P(A_{1:\Psi})$ of the choice of an activity pattern $A_{1:\Psi}$ and a model $P(x|A_{1:\Psi})$ of the choice of destination x conditional on the activity pattern $P(A_{1:\Psi})$ (Eq. 4.2). Here, an activity pattern $A_{1:\Psi} = (A_1, ..., A_\psi, ..., A_\Psi)$ is a sequence of activities $A_\psi = (\mathscr{A}_k, t^-, t^+)$ defined as an activity type

$\mathscr{A}_k$ and start and end times.

$$P_i(\hat{m}_{1:J}) = \sum_{a_{1:\Psi} \in \mathscr{C}} P(\hat{m}_{1:J} | a_{1:\Psi}) \cdot P(a_{1:\Psi}) \tag{4.1}$$

$$= \sum_{a_{1:\Psi} \in \mathscr{C}} \prod_{j=1}^{J} P(\hat{x}_j | x_\psi^j) \cdot P(A_{1:\Psi}) \cdot \prod_{\psi=1}^{\Psi} P(x_\psi | A_{1:\Psi}) \tag{4.2}$$

Figure 4.1 illustrates this decomposition in a train station. Assume a measurement $\hat{m}$ at an equal distance d from three points of interest in the domain of data relevance (DDR, see Bierlaire and Frejinger; 2008 and Ch. 3): two cafés, A and B, and a platform, platform 1. The probability of generating this measurement is in this case:

$$P(\hat{m}) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{d^2}{2\sigma^2}\right)\Big(P(\text{Café}) \cdot \Big(P(\text{Café A}|\text{Café}) + P(\text{Café B}|\text{Café})\Big) + P(\text{Platform}) \cdot P(\text{Platform 1}|\text{Platform})\Big)$$

assuming that the errors in latitude and longitude are independently and normally distributed (see Ch. 3).

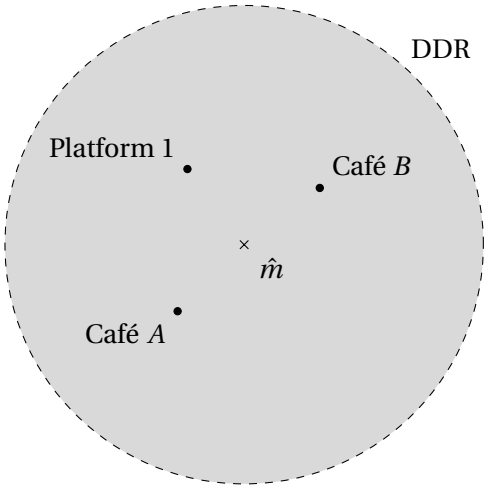


Figure 4.1 – A measurement $\hat{m}$ at an equal distance from three points of interest in the domain of data relevance (DDR): two cafés, A and B, and a platform, platform 1.

We present the choice of location for a given activity type later in this dissertation, in Ch. 5. Here, in this chapter, we propose a model for the choice of an activity-episode sequence.

We simultaneously model the choice of activity types, order, start times and durations of activity episodes in a sequence. The activity-episode sequence is modeled as a path in an activity network defining the activity type, duration and time of day. We develop a framework for choice set generation. The large dimensionality of the choice set is managed through a strategic sampling using a Metropolis-Hastings algorithm.

The activity-based approach is motivated by the fact that the choice of an activity pattern triggers the choice of locations (e.g. Bhat and Singh; 2000; Bowman and Ben-Akiva; 2001; Bierlaire and Robin; 2009). It is relevant for transportation policy simulation such as congestion pricing, toll lanes, and changing schedules for work or shops (Davidson et al.; 2007). It is also relevant in models of pedestrian movement (Papadimitriou et al.; 2009). The impact of timetables and platform allocations have been identified as a major challenge in pedestrian facilities such as train stations (Daamen; 2004, ch. 2). The impact of changes in train schedules and ticket purchase needs in a station is exemplified in Fig. 4.2.

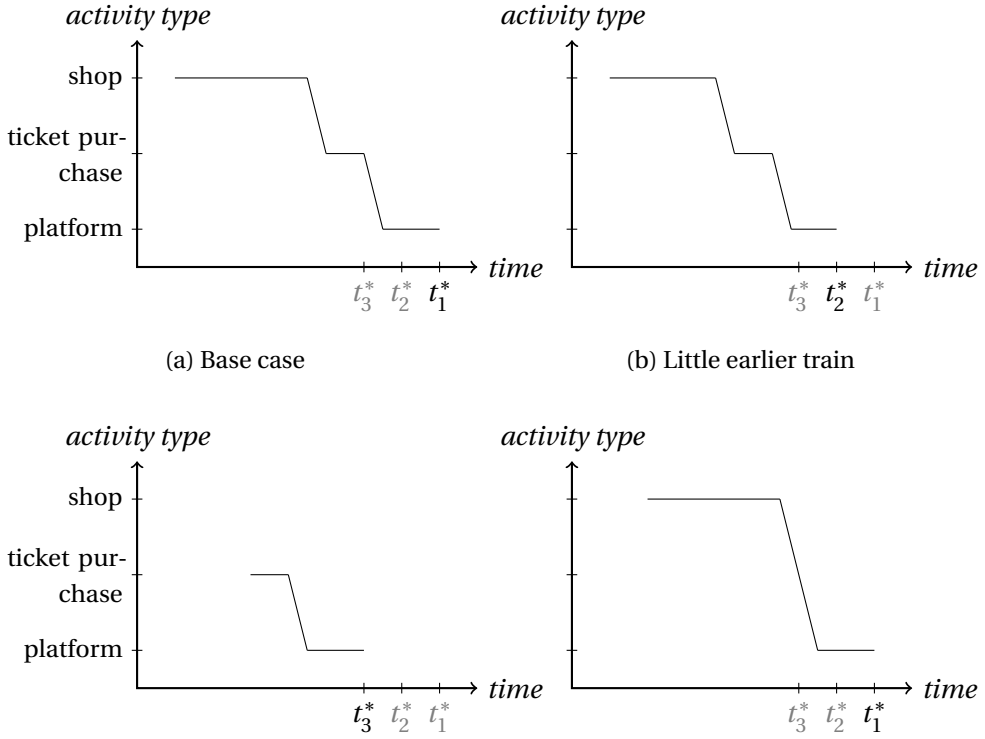
Several models accounting for interactions that shape participation in different activities have been proposed (see Section 2.3). Activity scheduling models over an entire-day framework are a mixture of rule-based algorithms, duration models and discrete choice frameworks. The main drawback of most of these models is the postulated rules: they are structured on home and tours from home, with models applied sequentially according to priorities of activity types. Very often, the large dimensionality of the problem (activity types, continuous time, number of episodes in the day) implies aggregation or hierarchy of dimensions (broad periods of time, mandatory vs non mandatory, primary vs secondary).

Our modeling approach is not tour-based and do not assume any priorities between activities. It can be applied to weekdays, weekends in urban contexts, or activities in a pedestrian facility, such as an airport or a supermarket. The chosen alternative is one sequence of activity episodes; utility is associated with the full pattern. Represented as a path in a network, the sequence is a single choice, contrary to Pinjari and Bhat (2010) who consider the activity-episode sequence as multiple choices of activity types and duration.

The methodology is developed in Section 4.2 and exemplified with WiFi traces on EPFL campus in Section 4.3.

4.2 Methodology

We present the concepts of activity network and activity path in Section 4.2.1 and the choice set generation using importance sampling in Section 4.2.2, with sampling correction of the utility in Section 4.2.3.



(a) Base case

(b) Little earlier train

(c) Much earlier train

(d) No ticket needed

Figure 4.2 – Adjustments of the activity-episode sequence and activity-episode durations to a modification in train schedules and ticket purchase needs. Fig. 4.2a represents the base case where the individual goes for shopping and then to buy a ticket, and finally on platform to take the train scheduled at t_1^*. Fig 4.2b represents a small modification of the train schedule from t_1^* to t_2^*: the individual still has time to go shopping and modifies the time spent for shopping and waiting on the platform accordingly to the time budget defined by the schedule. In Fig. 4.2c, the schedule is modified again, from t_2^* to t_3^*, and the individual does not have enough time for shopping: this activity episode is canceled. The need to buy a ticket modifies the activity-episode sequence in Fig 4.2d: the individual arrives later in the train station.

4.2.1 Representation of activity patterns: activity network and path

An activity network represents the choice set and contains all possible activity patterns. It is discrete with respect to activity types $\mathscr{A}_1, \mathscr{A}_2, ..., \mathscr{A}_K$ and time $\tau \in 1, 2, ..., T$. It is composed of links and nodes. Nodes $\mathscr{A}_{k,\tau}$ represent the performance by the individual of an activity type k for a unit of time τ. At a given unit of time τ, the number of nodes represent the available activity types K. There are two special nodes, start node s and end node e. They represent the beginning and the end of the observed activity pattern. In total, the activity network contains $KT + 2$ nodes. Edges connect some nodes and represent the fact that they are successively performed. s is connected to all nodes at the first time unit. All nodes at the last time unit are connected to e. All nodes of a given time unit t are connected with the nodes corresponding to the next time unit $t + 1$; it represents the choice of changing activity type or maintaining the activity type for one more time unit. In total, the activity network contains $2K + K^2(T - 1)$ edges.

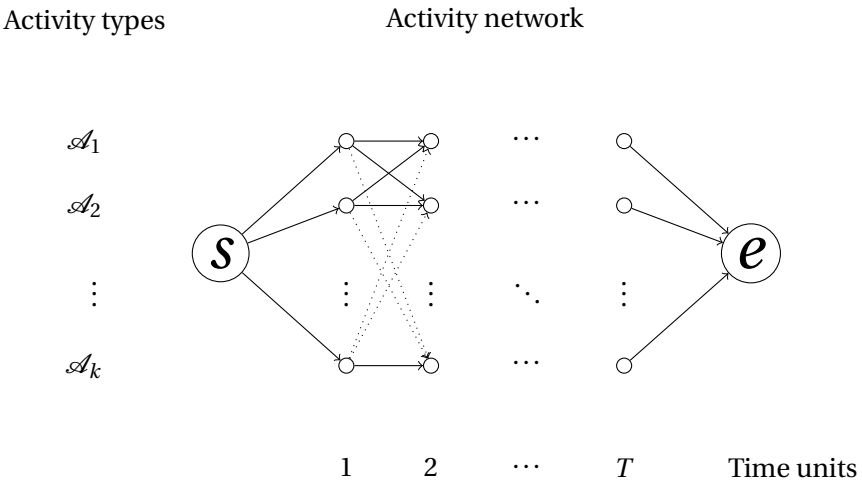


Figure 4.3 – The activity network

The activity network is a representation of the universal choice set. The first and last time units represent the period of time under observation (e.g., 4am and midnight for a day).

Activity paths $\mathscr{A}_{1:T}$ are the representation of activity patterns in an activity network. Activity paths are the alternatives of the choice process.

4.2.2 Choice set generation

The universal choice set contains K^T alternatives. With a large temporal resolution T, a complete enumeration of all paths and the estimation of a choice model with the universal choice set is infeasible. Different sampling strategies exist. Simple random sampling (SRS) is technically possible but the generated activity path are dominated alternatives and the model overfits the generated choice set (see Section 4.3).

Alternatively, importance sampling allows to estimate discrete choice models in the case of an extremely large choice set by including in the choice set alternatives that are more relevant. The sampling probabilities used for importance sampling must be known in order to get consistent estimators. With a very large number of elements, normalization is computationally impossible, since it would require the enumeration of all the possible paths. In the following subsections, we propose to generate a choice set for importance sampling using a Metropolis-Hastings algorithm for the sampling of paths, where the sampling strategy is defined by the utility ("strategic sampling").

The Metropolis-Hastings sampling of paths and the strategic sampling are described below. The utility function used to correct for importance sampling is later described in Section 4.2.3.

Metropolis-Hastings sampling of paths

The Metropolis-Hastings algorithm defined by Flötteröd and Bierlaire (2013) samples paths according to a given distribution. For the sampling of paths, the Metropolis-Hastings algorithm does not require a normalized distribution. It only requires an unnormalized version of the distribution, that we call *target weight* in the following discussion. A Markov chain with a predefined stationary distribution is generated by randomly modifying paths between an origin and a destination. Splice and shuffle operations on the paths (see Fig. 4.4) are made such that the paths appear with the frequency specified by the target weights.

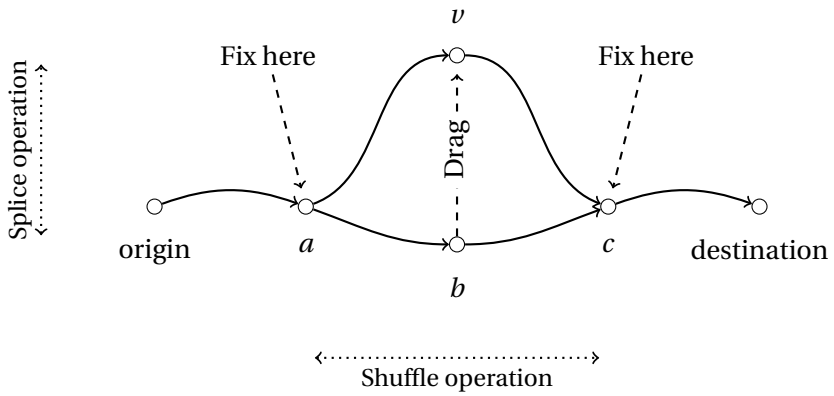


Figure 4.4 – Schematic figure of the splice and shuffle operations of the Metropolis-Hastings algorithm for sampling paths by Flötteröd and Bierlaire (2013). A state is a path Γ and three nodes a, b, c on this path. The "shuffle" operation maintains Γ but randomly modifies the three nodes a, b, c on the path. The "splice" operation randomly replaces the middle node b of the state by a new one v and connect nodes a and c with this newly inserted node v using shortest paths.

Metropolis-Hastings algorithms require the definition of a state variable, of a target distribution, and of a proposal distribution. In this category of algorithms, a proposal transition from one state to another is drawn from the proposal distribution and then accepted with a certain probability in order to reach a stationary distribution defined by the unnormalized

target distribution. Here, the state is defined as a path Γ and three nodes a, b, c on it.

The proposal distribution is a combination of the insertion probability $P_{\text{insert}}(v)$ and of the "shuffle" distribution P_{shuffle}. The insertion probability is defined as

$$P_{\text{insert}}(v) = \frac{e^{-\tilde{\mu}\left(\delta_{SP}(\text{origin},v)+\delta_{SP}(v,\text{destination})\right)}}{\sum_{w\in\mathcal{N}} e^{-\tilde{\mu}\left(\delta_{SP}(\text{origin},w)+\delta_{SP}(w,\text{destination})\right)}} \tag{4.3}$$

where $\tilde{\mu}$ is a scale parameter, δ_{SP} is the shortest path cost, and $\mathcal{N}$ is the set of all nodes. Shortest paths are used in the insertion probability for efficiency. Moreover, this probability is state independent and can be computed once for all nodes and does not need to be computed for each new state. Shortest path efficiency requires a link-additive cost definition for the proposal distribution. The "shuffle" operations leave the path unaffected and uniformly distribute a, b and c along the path.

The target weight does not need to be link-additive and can be defined by attributes related to the full path, such as primary activity (i.e., majority activity type) in the activity path. The target weight is defined as an exponentially decreasing function $e^{-\mu\delta(\Gamma)}$, where μ is a scale parameter and $\delta(\Gamma)$ a cost function for the path Γ. Both scale parameters μ and $\tilde{\mu}$ in target weight and proposal distribution, respectively, represent the variability of paths: $\mu, \tilde{\mu} = 0$ corresponds to uniform probability or weights, while $\mu, \tilde{\mu} \to 0$ does not consider anything else than the shortest path or the path with highest weight. For more details about the Metropolis-Hastings sampling of paths, see Flötteröd and Bierlaire (2013).

The main challenge in setting up the Metropolis-Hastings algorithm comes from the proper definition of the target weight (and a corresponding proposal distribution). As explained in Frejinger and Bierlaire (2010), "*the sample should include attractive alternatives*" in order to provide efficient estimators. In the context of activity choice modeling, a proper definition of target weight is not straightforward and cannot be defined from *a priori* knowledge about behavior. Moreover, once a target weight is defined, a corresponding proposal distribution must be found. If the proposal distribution does not vary enough and is concentrated around the shortest path, similar paths are generated and many draws are required for covering the relevant part of the state space according to the target weight; on the other hand, if the proposal distribution varies too much, many paths are irrelevant according to the target weight and are rejected.

Strategic sampling

We propose to use the utility function from a previously estimated model as target weight, similarly to what Lemp and Kockelman (2012) did on synthetic data. In this way, no assumption about the utility structure is made and we let the data speak. Regarding the proposal distribution, it must approximate the target distribution and be node-additive. We propose to estimate a model only with the node-additive attributes, i.e., with the time-of-day attributes,

on the same data and choice set than the initial model. The cost of each node, used in the shortest path computation, is the additive inverse of the utility and the insertion probability (Eq. 4.3) becomes the logit probability of the node-additive model.

In the estimation of both the initial model for target weights and the time-of-day model for the proposal distribution, the scale parameters are fixed to 1 for identification purpose. In the application of the model for strategic sampling in the Metropolis-Hastings algorithm, we fix $\mu = \tilde{\mu} = 1$.

4.2.3 Sampling correction in the utility

We assume the choice set to be the universal choice set containing all possible paths between s and e in the activity network. The sampling strategy for choice set generation presented in Section 4.2.2 requires the deterministic part of the utility to be corrected in order to estimate unbiased parameters (McFadden; 1978). According to Frejinger et al. (2009), a sampling correction term $\ln \frac{k_{\Gamma n}}{q(\Gamma)}$ must be added to the utility function for activity path Γ, where $k_{\Gamma n}$ is the number of times activity path Γ is drawn in $\mathscr{C}_i$ and $q(\Gamma)$ is the sampling probability of path Γ.

The sampling probability $q(\Gamma)$ is available using the unnormalized target weights $b(\Gamma)$ but require full enumeration for normalization: $q(\Gamma) = \frac{b(\Gamma)}{\sum_{\Gamma' \in \mathscr{U}} b(\Gamma')}$. In practice, the normalizing sum cancels out in the logit formulation and $b(\Gamma)$ can be used instead of $q(\Gamma)$.

The utility usually includes the time of day preference, the satiation effect and the schedule delay, as described in Section 2.3.5. In this case, the path utility depends on the utilities of individual nodes $\mathscr{A}_{k,\tau}$ and on the utilities of the activity episodes a of the activity path. The utility $V(\mathscr{A}_{k,\tau})$ of a node $\mathscr{A}_{k,\tau}$ represents the individual marginal utility from allocating one time unit to a certain activity type. It corresponds to the time-of-day utility and depends on both the activity type k and the time interval τ. It can be generally expressed as $\beta_{k,\tau} I_{k,\tau}$, where $I_{k,\tau}$ is a dummy variable (with value 1 if the activity path include node $\mathscr{A}_{k,\tau}$ and 0 otherwise) and $\beta_{k,\tau}$ is the corresponding parameter. In practice, some β's might be equal. The utility $V(a)$ of an activity episode a includes the satiation effect and the schedule delay.

Our modeling framework also gives the opportunity to add attributes that are not link-additive nor related to activity episodes, such as the repetition of certain activity types in the activity path, the structure of the activity path, the primary activity type, etc. Parameters can be interacted with socioeconomic variables.

The deterministic part of the utility correcting for the sampling of alternatives is:

$$\mu\left(\sum_{k=1}^{K}\sum_{\tau=1}^{T} V(\mathscr{A}_{k,\tau}) + \sum_{a \in \mathscr{A}_{1:T}} V(a) + V(\Gamma)\right) + \ln \frac{k_{\Gamma n}}{b(\Gamma)}. \tag{4.4}$$

Concrete instances of node utility $V(\mathscr{A}_{k,\tau})$, activity-episode utility $V(a)$ and activity path utility $V(\Gamma)$ are presented in the case study below.

4.3 Pedestrian case study on EPFL campus

The choice of an activity pattern also triggers people's behavior in pedestrian facilities (Bierlaire and Robin; 2009). Demand management measures can be applied in pedestrian facilities, such as changing schedules (train in train stations, planes in airports, concerts in music festivals or class schedules in universities). Our model is particularly suitable for pedestrian facilities, since it does not assume a tour structure nor a "home" activity type.

In particular, transport hubs (e.g., train stations or airports) are key nodes of a multimodal transport system (with buses, metro, car and bike sharing). Train stations are located in the city centers and include shops and services. All these activities combined with the growth in the number of passengers increase pedestrian flows and threaten the functioning of the train station. Understanding demand for activities is of utmost importance to define appropriate planning policies.

As a proof of concept, we apply our methodology to WiFi traces collected on the EPFL campus. EPFL campus approximately hosts 13'000 people per day. Similarly to transport hubs, some of them follow schedules (class schedules instead of train schedules) and perform several different activities, such as going to class, having lunch, etc.

Section 4.3.1 describes the data used in this case study, Section 4.3.2 describes the choice set generation process and the choice model, and results are presented in Section 4.3.3.

4.3.1 Data source and activity network

We collected data as defined in Ch. 3. Campus users authenticate themselves on the WiFi network through WPA using a Radius server. Accounting is one of the process on the Radius server. It allows to associate a MAC address with a username. Each measurement was associated with a unique identifier and a category, such as *employee* or *civil engineering student, bachelor*. Data were then anonymized by deleting the MAC address. Details about the data collection campaign and data cleaning can be found in Appendix A.2 and raw data are available in Danalet (2015).

The Bayesian approach described in Ch. 3 was then applied to the raw WiFi traces in order to detect activity-episode sequences. It merges WiFi traces with data from the map of the campus, measures of attractivity of each destination and time constraints. The precise definition of these data, and in particular of attractivity measures, can be found in Appendix A.2.

We assume 8 activity types: classrooms, shops, offices, restaurant, library, lab, other and not being detected. The types "Office", "Classroom" and "Lab" are based on norm DIN 277 defined

by the *Deutsches Intitut für Normung*. The types "Shops", "Restaurant", "Library" and "Other" are extracted from a list of points of interest from http://map.epfl.ch.

There are $T = 12$ time units in the activity network, from 7am to 7pm, at each hour.

4.3.2 Choice set and choice model

The full choice set cannot be enumerated (8^{12} paths) in this case study. As described in Section 4.2.2, the Metropolis-Hastings algorithm uses the utility of a first model estimated from a choice set generated with simple random sampling (Table 4.1) as target weights and a time-of-day, node-additive model (Table 4.2) as proposal distribution. We adapted the Java code[1] from Flötteröd and Bierlaire (2013) to our needs (in particular non-link additive weights), with $P_{\text{splice}} = 0.75$.

Given the randomly generated choice set, the first model assigns a probability of almost 1 to the chosen alternative, and a probability of almost 0 for the other elements of the choice set. It results in a final log-likelihood of almost 0 (-47.218). Extra attributes cannot be added in this specification, identified and be significant because the log-likelihood is flat, given all other attributes.

The deterministic part of the utility used for the estimation of this first model is an instance of Eq. 4.4. Node utility $V(\mathscr{A}_{k,\tau})$, activity-episode utility $V(a)$ and activity path utility $V(\Gamma)$ are defined as follows:

$$V(\mathscr{A}_{k,\tau}) = \beta_{\mathscr{A}_k \tau,\,\text{group}} \mathbb{1}_{k,\tau} \mathbb{1}_{\text{group}} \tag{4.5}$$

$$V(a) = \eta_{\mathscr{A}_k} \ln(|a|) \mathbb{1}_{A(a)=\mathscr{A}_k} \tag{4.6}$$

$$V(\Gamma) = \beta_{|\Gamma|_{\mathscr{A}_k}} \mathbb{1}_{|\Gamma|_{\mathscr{A}_k}} \tag{4.7}$$

where $|a|$ is the duration of activity episode a, $A(a)$ is the activity type of activity episode a and $|\Gamma|_k$ is the number of activity episodes in activity path Γ with activity type k. The first 14 attributes $\beta_{\mathscr{A}_k \tau,\,\text{group}}$ in Table 4.1 represent time-of-day preferences, with $\mathscr{A}_k \in$ {lab, library, office, restaurant, shop, NA}, $\tau \in \{7, \ldots, 18\}$ and group $\in$ {students, employees}. The different η's represent the satiation parameters. They multiply the logarithm of the duration of the activity episode a and are specific to an activity type ($\mathbb{1}_{A(a)=\mathscr{A}_k}$ has value 1 if activity episode a corresponds to activity type $\mathscr{A}_k$ and 0 otherwise). Finally, the different $\beta_{|\Gamma|_{\mathscr{A}_k}}$ represent a preference for a given number of activity episodes in the full activity path for different activity types. Specifically, $\beta_{3+\text{ lab episodes}}$, $\beta_{3+\text{ resto episodes}}$, $\beta_{2\text{ NA episodes}}$ and $\beta_{3\text{ NA episodes}}$ represent the preference for 3 or more activity episodes being "Lab", 3 or more being "Restaurant" and 2 being "not being detected", respectively.

The time-of-day preferences have expected signs in Table 4.1. Employees work in labs

[1] Available on http://people.kth.se/~gunnarfl/bioroute.html.

Description	Coeff. estimate	Robust Asympt. std. error	t-stat
$\beta_{\text{lab 14-19, students}}$	-3.12	0.563	-5.55
$\beta_{\text{lab 12-14, students}}$	-9.17	1.39	-6.59
$\beta_{\text{lab 7-12, 14-19, employees}}$	1.65	0.271	6.08
$\beta_{\text{library 7-12, employees}}$	-2.08	0.422	-4.93
$\beta_{\text{office 7-12, 14-19, employees}}$	1.69	0.393	4.30
$\beta_{\text{restaurant 12-14, employees}}$	1.22	0.502	2.43
$\beta_{\text{restaurant 7-12, 14-19, employees}}$	1.51	0.249	6.06
$\beta_{\text{shop 12-14, students}}$	-7.36	1.24	-5.92
$\beta_{\text{shop 7-12, 14-19, students}}$	-1.16	0.538	-2.16
$\beta_{\text{NA 7-8, students}}$	4.27	0.995	4.29
$\beta_{\text{NA 8-12, students}}$	1.40	0.498	2.82
$\beta_{\text{NA 17-19, students}}$	1.75	0.568	3.08
$\beta_{\text{NA 9-17, employees}}$	1.43	0.296	4.84
$\beta_{\text{NA 7-9, 17-19, employees}}$	3.34	0.554	6.02
$\eta_{\text{Office, Lab, Classroom}}$	5.22	0.764	6.83
$\eta_{\text{Restaurant, Library, Other}}$	7.85	1.11	7.10
η_{Shop}	7.33	0.894	8.20
η_{NA}	2.75	0.393	7.00
$\beta_{\text{3+ lab episodes}}$	-5.03	0.952	-5.28
$\beta_{\text{3+ resto episodes}}$	-2.50	0.759	-3.29
$\beta_{\text{2 NA episodes}}$	5.09	1.01	5.02
$\beta_{\text{3 NA episodes}}$	3.71	1.17	3.16

Number of observations = 1087
Number of estimated parameters = 22
$\mathscr{L}(\beta_0) = -5016.636$
$\mathscr{L}(\hat{\beta}) = -47.218$
$\rho^2 = 0.991$
$\bar{\rho}^2 = 0.986$

Table 4.1 – First model, estimated using simple random sampling. It is used as target weight in the Metropolis-Hastings algorithm.

($\beta_{\text{lab 7-12, 14-19, employees}} > 0$), while students don't ($\beta_{\text{lab 14-19, students}}$, $\beta_{\text{lab 12-14, students}} < 0$). The library does not attract employees ($\beta_{\text{library 7-12, employees}}$<0), because they are often in their office ($\beta_{\text{office 7-12, 14-19, employees}} > 0$) or in restaurants ($\beta_{\text{restaurant, employees}}$'s > 0). Students are not often shopping ($\beta_{\text{shop 12-14, students}}$, $\beta_{\text{shop 7-12, 14-19, students}} < 0$). Finally, people are likely not to be on campus at any time of the day, more in the early morning for students (between 7:00 and 8:00, before classes start) and more outside office hours for employees (office hours are defined here as from 9:00 to 17:00). The satiation parameters are positive, expressing the preference for longer activity episodes as compared to the randomly generated activity paths in the choice set.

The time-of-day, node-additive model presented in Table 4.2 corresponds to a deterministic part of the utility containing only elements as in Eq. 4.5.

Each successive state of the Metropolis-Hastings algorithm is very likely to be similar to the previous one. This similarity stabilizes with the distance d between iterations. Stabilization means independence of the sampled paths. A similarity measure is defined in Flötteröd and Bierlaire (2013):

$$\phi(d) = \frac{1}{K}\sum_{\kappa=1}^{K} \frac{|\Gamma^k \cap \Gamma^{k+d}|}{\frac{1}{2}(|\Gamma^k| + |\Gamma^{k+d}|)} \tag{4.8}$$

where $|\Gamma^k \cap \Gamma^{k+d}|$ is the number of identical nodes in the paths generated in iterations k and $k+d$. Fig. 4.5 shows that the similarity stabilizes at 0.47 after a warming-up period. Similarly to Flötteröd and Bierlaire (2013), we fit a linear regression model on 10 consecutive similarity values $\phi(d), ..., \phi(d+9)$ for consecutive distances $d, ..., d+9$. We assume the sampled paths to be independent when the absolute slope of the linear model is below 10^{-3}. Distance to reach independence with the utility function for employees is presented in Fig. 4.5 for $\mu = 1.0$.

We estimate a logit model using a choice set containing 100 activity paths for each observation based on the Metropolis-Hastings sampling of paths, sampling one path every d iterations, with $\mu = 1$. The generated paths are assumed to be independent from each other but are still similar to each other. Out of the 173′400 (100 × 1734 observations) generated paths, 100′776 are different from each other. Sampling correction is included in the utility as described in Section 4.2.3.

4.3.3 Estimation results

A model is estimated with choice sets of 100 alternatives generated with the Metropolis-Hastings sampling of paths. For comparison between simple random sampling and strategic sampling, we use the specification of the model used as target weight in the Metropolis-Hastings algorithm (Fig. 4.1). The final log-likelihood using strategic sampling is -1044.266, lower than with simple random sampling (-47.218). It shows that strategic sampling allows to decrease the final log likelihood and thus to increase the number of explanatory variables for

Description	Coeff. estimate	Robust Asympt. std. error	t-stat
$\beta_{\text{NA, 17-19, employees}}$	0.263	0.0302	8.70
$\beta_{\text{NA, 14-17, students}}$	-0.222	0.191	-1.16
$\beta_{\text{NA, 7-8, students}}$	0.349	0.0281	12.44
$\beta_{\text{NA, 7-9, employees}}$	0.326	0.0262	12.43
$\beta_{\text{NA, 17-19, students}}$	1.14	0.187	6.09
$\beta_{\text{NA, 12-14, students}}$	0.333	0.247	1.35
$\beta_{\text{NA, 8-12, students}}$	0.141	0.0180	7.84
$\beta_{\text{NA, 9-17, employees}}$	1.11	0.213	5.20
$\beta_{\text{classroom, 14-19, employees}}$	0.335	0.294	1.14
$\beta_{\text{classroom, 14-19, students}}$	-0.351	0.183	-1.92
$\beta_{\text{classroom, 12-14, employees}}$	0.00915	0.460	0.02
$\beta_{\text{classroom, 12-14, students}}$	-0.336	0.337	-1.00
$\beta_{\text{classroom, 7-12, employees}}$	-0.723	0.397	-1.82
$\beta_{\text{classroom, 7-12, students}}$	0.598	0.262	2.28
$\beta_{\text{lab, 14-19, employees}}$	1.05	0.248	4.21
$\beta_{\text{lab, 14-19, students}}$	-2.97	0.806	-3.68
$\beta_{\text{lab, 12-14, employees}}$	0.915	0.264	3.47
$\beta_{\text{lab, 12-14, students}}$	-3.40	1.09	-3.13
$\beta_{\text{lab, 7-12, employees}}$	1.01	0.229	4.40
$\beta_{\text{library, 14-19, employees}}$	-0.624	0.553	-1.13
$\beta_{\text{library, 14-19, students}}$	-0.848	0.235	-3.61
$\beta_{\text{library, 12-14, employees}}$	-0.575	0.481	-1.20
$\beta_{\text{library, 12-14, students}}$	-0.859	0.345	-2.49
$\beta_{\text{library, 7-12, employees}}$	-1.57	0.508	-3.09
$\beta_{\text{library, 7-12, students}}$	-0.0229	0.293	-0.08
$\beta_{\text{office, 14-19, employees}}$	1.41	0.246	5.73
$\beta_{\text{office, 14-19, students}}$	0.0890	0.132	0.67
$\beta_{\text{office, 12-14, employees}}$	1.40	0.249	5.62
$\beta_{\text{office, 12-14, students}}$	0.501	0.290	1.73
$\beta_{\text{office, 7-12, employees}}$	1.12	0.228	4.92
$\beta_{\text{office, 7-12, students}}$	0.708	0.216	3.27
$\beta_{\text{restaurant, 14-19, employees}}$	0.920	0.255	3.60
$\beta_{\text{restaurant, 14-19, students}}$	-0.410	0.185	-2.21
$\beta_{\text{restaurant, 12-14, employees}}$	0.136	0.0259	5.26
$\beta_{\text{restaurant, 12-14, students}}$	0.665	0.286	2.32
$\beta_{\text{restaurant, 7-12, employees}}$	0.563	0.218	2.58
$\beta_{\text{restaurant, 7-12, students}}$	-0.151	0.267	-0.57
$\beta_{\text{shop, 14-19, employees}}$	-0.194	0.342	-0.57
$\beta_{\text{shop, 14-19, students}}$	-2.13	0.280	-7.64
$\beta_{\text{shop, 12-14, employees}}$	-0.0553	0.472	-0.12
$\beta_{\text{shop, 12-14, students}}$	-2.65	0.810	-3.27
$\beta_{\text{shop, 7-12, employees}}$	-0.473	0.354	-1.34
$\beta_{\text{shop, 7-12, students}}$	-1.83	0.594	-3.09

Number of observations = 1087
Number of estimated parameters = 43
$\mathscr{L}(\beta_0) = -5016.636$
$\mathscr{L}(\hat{\beta}) = -453.225$
$\rho^2 = 0.910$
$\bar{\rho}^2 = 0.901$

Table 4.2 – The time-of-day, node-additive model used as proposal distribution in the Metropolis-Hastings algorithm. The shortest path for connecting the insertion node in the splice operation of the Metropolis-Hastings algorithm is based on a time-of-day model, as described in Section 4.2.2.

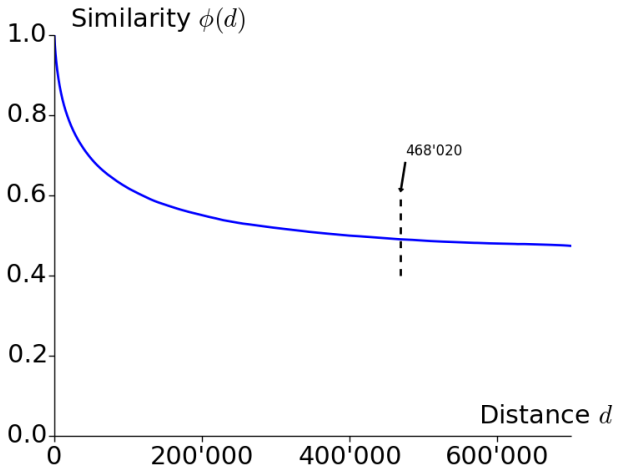


Figure 4.5 – Similarity measure as a function of the distance. 10^6 paths were generated using strategic sampling, with $\mu = 1.0$. Distance to reach independence is represented by the dashed line, with its numerical value.

the choice of activity type, duration, time of the day and order of the episodes. Table 4.3 shows the estimated parameters for a model using strategic sampling with $\mu = 1$. Compared to the initial model with simple random sampling (Table 4.1), the number of estimated parameters significantly different from zero increases from 22 to 39.

The final log-likelihood $\mathscr{L}(\hat{\beta})$ is not close to zero anymore (-400.633) with the new model (Table 4.3) and the adjusted rho-square $\bar{\rho}^2$ is not close to 1 (0.912), compared to the model estimated with simple random sampling (Table 4.1). It indicates that the sampled choice sets are not dominated anymore by the chosen alternative. The current choice set is not fully dominated by the observed choice, given the choice model. Extra variables, such as schedule delay, have not been included simply because they were not significantly different from zero.

4.3.4 Validation

Models obtained through simple random sampling (Table 4.1) and strategic sampling (Table 4.3) are validated by estimating each model on the observations corresponding to a random selection of 80 % of the individuals and applying the model with the estimated parameters to the observations of the remaining 20 % of the individuals. The predicted probabilities for the chosen alternative are presented in Fig. 4.6.

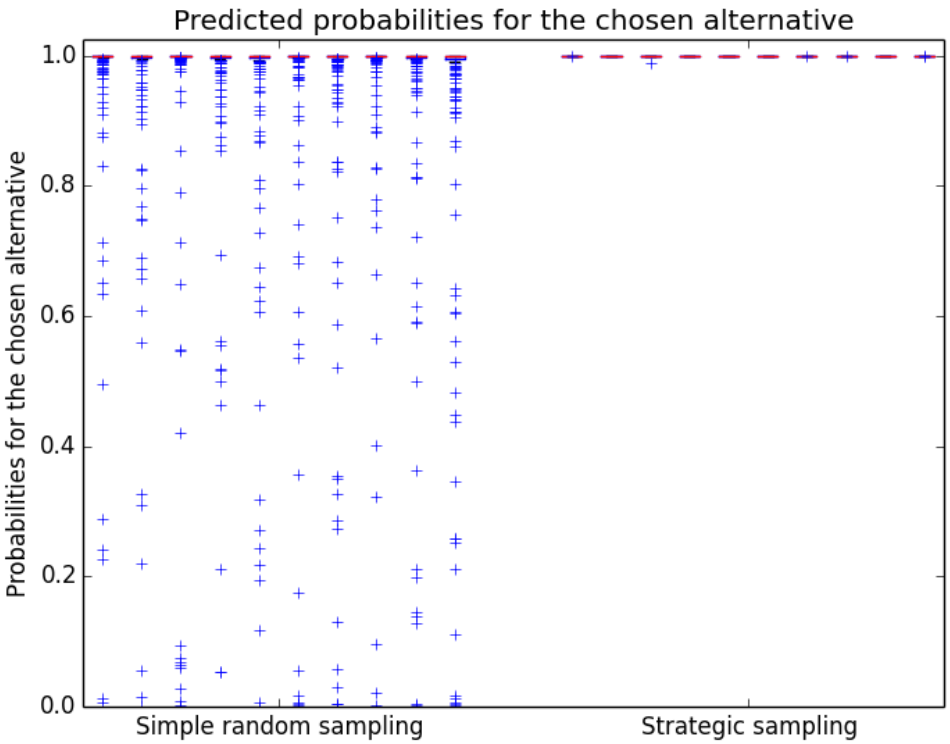


Figure 4.6 – Boxplot of the predicted probabilities for the chosen alternative, both using simple random sampling and strategic sampling with 10 elements in the choice set and 10 replications each ($\mu = 1$). The blue crosses are outliers. The mean and the quartiles are almost 1 in both cases.

Description	Coeff. estimate	Robust Asympt. std. error	t-stat
$\beta_{\text{NA 7-12, students}}$	2.33	0.285	8.17
$\beta_{\text{NA 12-14, students}}$	2.76	0.497	5.54
$\beta_{\text{NA 14-17, students}}$	2.90	0.392	7.41
$\beta_{\text{NA 17-19, students}}$	2.83	0.343	8.24
$\beta_{\text{NA 7-9, 17-19, employees}}$	2.91	0.303	9.60
$\beta_{\text{NA 9-17, employees}}$	1.96	0.251	7.81
$\beta_{\text{classroom 14-19, employees}}$	1.71	0.312	5.47
$\beta_{\text{classroom 7-12, students}}$	0.478	0.238	2.01
$\beta_{\text{lab 14-19, employees}}$	1.46	0.158	9.19
$\beta_{\text{lab 7-12, employees}}$	1.22	0.152	8.02
$\beta_{\text{office 7-12, employees}}$	-0.269	0.107	-2.51
$\beta_{\text{other 7-19, employees}}$	-0.699	0.168	-4.17
$\beta_{\text{restaurant 12, students}}$	2.69	0.527	5.10
$\beta_{\text{restaurant 12-14, employees}}$	-0.540	0.138	-3.93
$\beta_{\text{shop 14-19, employees}}$	1.46	0.343	4.27
$\beta_{\text{shop 12-14, students}}$	1.17	0.114	10.30
$\beta_{\text{shop 7-12, employees}}$	1.10	0.330	3.32
$\eta_{\text{office, lab, classroom}}$	-6.85	0.379	-18.09
$\eta_{\text{restaurant, library, other}}$	-6.58	0.360	-18.31
η_{shop}	-3.72	0.278	-13.40
η_{NA}	-7.63	0.541	-14.12
$\beta_{\text{0 NA episode}}$	10.8	1.34	8.08
$\beta_{\text{1 NA episode}}$	7.90	0.955	8.27
$\beta_{\text{2 NA episodes}}$	-2.22	0.840	-2.65
$\beta_{\text{3 NA episodes}}$	-5.32	0.874	-6.08
$\beta_{\text{0 classroom episode, employees}}$	10.3	0.887	11.65
$\beta_{\text{1 classroom episode, employees}}$	6.52	0.823	7.92
$\beta_{\text{1 classroom episode, students}}$	-0.840	0.370	-2.27
$\beta_{\text{0 lab episode, employees}}$	6.05	0.557	10.87
$\beta_{\text{0 lab episode, students}}$	7.22	0.748	9.65
$\beta_{\text{1 lab episode, employees}}$	2.17	0.363	5.98
$\beta_{\text{0 library episode, employees}}$	2.72	0.335	8.10
$\beta_{\text{0 library episode, students}}$	4.77	0.495	9.64
$\beta_{\text{0 office episode, employees}}$	4.05	0.422	9.59
$\beta_{\text{1 office episode, employees}}$	1.42	0.307	4.62
$\beta_{\text{0 restaurant episode}}$	4.11	0.365	11.28
$\beta_{\text{1 restaurant episode}}$	1.46	0.221	6.58
$\beta_{\text{1 shop episode}}$	-3.87	0.573	-6.76
$\beta_{\text{2+ shop episodes}}$	-3.49	1.08	-3.24

Number of observations = 1087
Number of estimated parameters = 39
$\mathscr{L}(\beta_0) = -5016.636$
$\mathscr{L}(\hat{\beta}) = -400.633$
$\rho^2 = 0.920$
$\bar{\rho}^2 = 0.912$

Table 4.3 – Model estimated using strategic sampling.

For simple random sampling and strategic sampling, the predicted probabilities for the chosen alternative are mostly close to one; it validates both approaches, since they both properly predict most of the choices. In the case of simple random sampling, alternatives in the choice set are purely random and not realistic. Consequently, the model (Table 4.1) is a crude approximation. It evaluates the tradeoffs between realistic alternatives (the chosen alternatives) and unrealistic alternatives (the choice set). Some outliers are not properly predicted (probabilities down to almost 0 for some of them). With strategic sampling, the model (Table 4.3) is improved and outliers in predictions for the 20 % of the population are systematically removed. It shows an improvement of the model by using importance sampling; it also means that the generated choice set using strategic sampling is not realistic enough to decrease the probability of the chosen alternative.

4.4 Conclusion

Our approach models the choice of activity sequences of individuals and evaluates the preferences for different activity types in time, the satiation effect for each activity type, primary activity and schedule delay effects. Our model is not home-based, nor tour-based. It can adapt to different contexts and activity types. The large dimensionality of the problem is managed through importance sampling techniques, using a Metropolis-Hastings sampling of path associated with strategic sampling.

An important feature of our approach is that it allows to add in the utility function variables that are not specific to time-of-day preferences or activity episodes, but are related to the path itself. Patterns (e.g., a office-restaurant-office pattern for employees for lunch) or primary activity can be included in the utility function for the day.

One key issue of this approach concerns the choice set. The choice set is extremely large and the universal choice set cannot be used. Consideration choice set is very difficult to define. Importance sampling seems the best strategy to define a choice set but needs proper weights. In this paper, we use strategic sampling, i.e., the choice probabilities of a first model as weights. This strategy works and allows to include more parameters in the model than simple random sampling. It has been validated and used for forecasting.

For estimation, we need to generate several elements in the choice set per observation. Many activity paths need to be generated and the distance between states to reach independence in the Metropolis-Hastings algorithm is in the order of $d = 10^5$ in our case study (Section 4.3.2). The computational burden was manageable in our case study. However, the Metropolis-Hastings sampling of paths was originally developed for route choice models and future research on its efficiency might improve its usability for large scale activity path choice models. We suggest to develop a Metropolis-Hastings sampling of activity paths that drags an activity episode instead of a point. The state space would consists of tuples $(\Gamma, a, b_{t^-}, b_{t^+}, c)$, where b_{t^-} and b_{t^+} are the new activity episode bounds. The splice operation generates a new activity episode between v_{t^-} and v_{t^+} for activity type k, as described in Fig. 4.7. This future algorithm

would match more closely the observed behavior in activity path choice set generation.

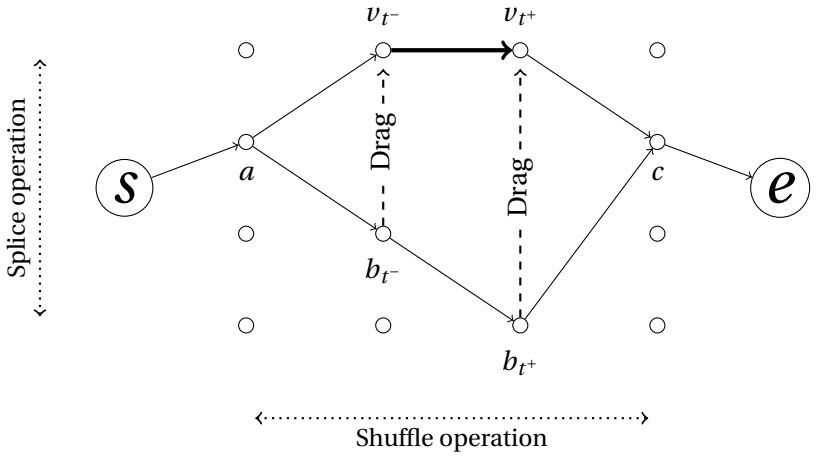


Figure 4.7 – Schematic figure of the splice and shuffle operations of the Metropolis-Hastings sampling of activity paths

5 Location choice with panel effect

In collaboration with Loïc Tinguely and Matthieu de Lapparent

5.1 Introduction

In this chapter, we explore the second modeling step presented in Fig. 1.1 and in Ch. 4.1. Once the activity path has been chosen (i.e., the activity type and the start and time of the episode, see Chapter 4), the location choice happens. In a train station, an example of decomposition of behavior would be an activity-episode sequence starting by *buying a ticket* at *a ticket machine*. Chapter 4 models when the individual is going to *buy a ticket*, before or after which other activities in the sequence. Here, we model at which *specific ticket machine* the individual is going to buy the ticket.

Destination choice models rely mostly on static frameworks with cross-sectional data, collected at one point in time (e.g., Ben-Akiva and Lerman; 1985; Zhu and Timmermans; 2011; Scott and He; 2012; Kalakou et al.; 2014). Panel data are difficult and expensive to collect (Yang and Timmermans; 2015), and sometimes inexistant, e.g., for the analysis of induced trafic at an aggregate level (Weis and Axhausen; 2009). In absence of actual panel data, pseudo panel data are constructed by grouping individuals from cross sectional sectional data into cohorts and by considering behavior of cohorts as individuals (Deaton; 1985; Weis and Axhausen; 2009; McDonald; 2015). However, actual panel data from technology-based data are more and more common in the literature (e.g., Carrion et al.; 2014; Kazagli et al.; 2014). Network traces (e.g., WiFi traces or cell tower data) are increasingly available and used for location choices (see Section 2.4). Compared to traditional surveys, network traces follow individuals on a longer period (see Section 2.2.1). Thus, it becomes possible to collect activity-episode sequences covering several days, weeks or months. Location choice models must adapt to these new data. This chapter specifically develops a modeling framework to account for panel data in location choices. It allows to understand people's habits in their decision process, while correcting for serial correlation.

We present the methodology in Section 5.2 and apply it to a pedestrian case study in Section 5.3, including validation and forecasting. We conclude in Section 5.4.

5.2 Methodology

We consider that an individual n repeatedly visits locations. For each individual n, we assume a sequence of events $\{1, ..., t_n, ..., T_n\}$. This sequence is exogenous and individual specific. At each event, a location choice is made. The indicator y_{int_n} is 1 if individual n selects location i for event t_n. The time interval between two events vary, as well as the number T_n of events per individual. To make the notation light, we use t instead of t_n in the following developments.

A sequence of events with varying time intervals between the decisions is typical for the choice of buying or selling for investors in the stock market (e.g., Robin and Bierlaire; 2012). It is also common when considering the activity location choice conditional on an activity type (e.g., Kalakou et al.; 2014; Ton; 2014). The modeling and forecasting of choices of activity type and time intervals between events is covered in Ch. 4.

We use a logit model for the choice of a location i. We consider an individual n and the sequence of its activities during a day: $a_{1:\Psi_n} = (a_1, a_2, ..., a_{\psi_n}, ..., a_{\Psi_n})$, where $a_{\psi_n} = (i, t^-, t^+)$ is an activity episode at location i with start time t^- and end time t^+. In this section, we present three models: a *static* model, a *dynamic model without agent effect* model and a *dynamic model with agent effect*.

We associate a utility U_{int} with a location i:

$$U_{int} = V_{int} + \varepsilon_{int} \tag{5.1}$$

where $i \in \mathscr{C}_{nt}$ and $\mathscr{C}_{nt}$ is the choice set of all available locations at time t for individual n. This model is simple to estimate when we assume that $\varepsilon_{int} \overset{iid}{\sim} EV(0,1)$ across i, n and t, i.e., a static logit model. It ignores two aspects: dynamics and serial correlation.

First, the choice at a certain time t may depend on previous choices. Individuals tend to have state dependence towards already visited locations. For simplification purpose, we make three additional assumptions. First, we assume a dynamic process of order one: the current level of utility of location i partly depends on the previously chosen location for the same type of activity. Second, the state dependence is location specific: utility for a location depends only on previous choice of this location. Third, we assume that the weight ρ of this state dependence is the same for every individuals n and every locations i (the assumption is restrictive and could be relaxed by considering variations across locations and individuals):

$$U_{int} = V_{int} + \rho y_{in(t-1)} + \varepsilon_{int} \tag{5.2}$$

where $y_{in(t-1)}$ is a dummy variable with value one if location i was chosen by individual n as the previous location in the previous activity episode with the same activity type, and 0

otherwise. The coefficient ρ measures the effect of previous experience of the location on its current utility. It can be interacted with the time of day, e.g., the choice of a catering location for a coffee break in the afternoon depends on the previous location choice for the same activity (catering) in the afternoon, ignoring the catering activity episodes for lunch in-between.

We assume that the time interval between two events does not change the impact of the previous experience, i.e., the memory of the previous activity location choice. The choice probability of an activity location is only influenced by a previous visit at the same activity location. Duration between two events does not affect choice probability.

We initially assume that the previous choice $y_{in(t-1)}$ is independent of the error term ε_{int} (strict exogeneity assumption) and that ε_{int} are independent and identically distributed across i, n and t. We term such a model a *dynamic model without agent effect*.

The error terms ε_{int} model the unobserved factors. In the *static* and the *dynamic model without agent effect*, we assume that they are independently distributed over time, individuals and locations. In practice, it is very likely that they share time-invariant components associated with the decision-maker, thereby generating *serial correlation*. This raises the second issue of the static model. For example, in the successive choice of a restaurant, taste for healthy food is usually unobserved (Burton et al.; 2014; Chen and Yang; 2014). In our context, it can be considered as an unobserved time-invariant factor[1].

As a consequence, the lagged variable $y_{in(t-1)}$ and the unobserved factors ε_{int} are correlated since they both depend on the time-invariant factor, also known as *agent effects*. This is called *endogeneity*. It has to be taken into account to avoid bias in the estimation of the parameters of the model.

We relax the independence assumption of error terms $\varepsilon_{in(t-1)}$ and ε_{int} by replacing the original single error term ε_{int} by the sum of two error terms: $\alpha_{in} + \varepsilon'_{int}$. α_{in} is the agent effect. It is time-invariant and represents the long-term preferences of individual n over time for location i. The agent effect α_{in} does not vary over time but varies across individuals (*inter-individuals variability*). ε'_{int} is the unobserved heterogeneity and represents the short-term variation of preferences of individual n (*intra-individual variability*). ε'_{int} are independent across time and individuals. The utility function becomes:

$$U_{int} = V_{int} + \rho y_{in(t-1)} + \alpha_{in} + \varepsilon'_{int}. \tag{5.3}$$

In classical dynamic panel data models with agent effects and lagged dependent variables, solving endogeneity bias in estimation by some maximum likelihood techniques requires the computation of the marginal/steady state choice probability for the first observed outcome of the dependent variable (Wooldridge; 2005, p.40: “use the joint distribution of all

[1] We agree that taste may change in the lifecycle of an individual, but not during the time horizon of the data we use for the application.

outcomes on the response—including that in the initial time period—conditional on unobserved heterogeneity and observed strictly exogenous explanatory variables"). This is often referred as the *initial conditions problem* in econometrics (Heckman; 1981; Hsaio; 2003; Train; 2003; Wooldridge; 2005). Computation of such marginal probability is intractable except for some simple binary models (see Bhargava and Sargan; 1983, Hsaio; 2003 (Section 4.3) and Wooldridge; 2005). Several authors have proposed circumventing strategies to solve this problem (see Hsaio; 2003 and Wooldridge; 2005 for reviews). We here build up on the Wooldridge (2005) correction method.

5.2.1 Correcting endogeneity for dynamic discrete choice models

In general, endogeneity must be corrected to get consistent estimates (Train; 2003, Ch. 13). *Control functions* capture the relationship between the unobserved factors and the observed variables and "absorb" endogeneity (Heckman; 1978).

Wooldridge (2005) proposes to model the distribution of the agent effect α_{in} conditional on the initial value and any exogenous explanatory variables:

$$\alpha_{in} = a + by_{in0} + c'\bar{x}_n + \xi_{in} \tag{5.4}$$

where ξ_{in} is normally distributed, $\xi_{in} \sim N(0, \Sigma_\alpha)$, with Σ_α is a matrix of parameters to be estimated[2], and $\bar{x}_n$ is a vector of time-invariant explanatory variables (i.e., long-term preferences, socioeconomic characteristics). The utility of the *dynamic model with agent effect* is:

$$U_{int} = V_{int} + \rho y_{in(t-1)} + a + by_{in0} + c'\bar{x}_n + \xi_{in} + \varepsilon'_{int}. \tag{5.5}$$

The endogeneity issue is addressed with this utility function, given that the assumption in Eq. 5.4 is valid (see Wooldridge (2005) for a detailed discussion). The contribution of a series of observations y_{int} at times $t = 1, ..., T$ for individual n to the likelihood function, conditional on the initial value y_{in0} and the agent effects $\alpha_n = \{\alpha_{in}, \forall i\}$, is:

$$P(y_{in1}, y_{in2}, ..., y_{int} | y_{in0}, \alpha_n) = \prod_{t=1}^{T} P(y_{int} | y_{in0}, y_{in(t-1)}, \alpha_n). \tag{5.6}$$

Note that we do not model the first choice y_{in0}. Given our assumptions, it turns out that our estimator is a conditional maximum likelihood estimator. It is asymptotically equivalent to the full information maximum likelihood estimator. Only efficiency is affected.

When integrating out the agent effects $\alpha_n \in \mathbb{R}^{dim(i)}$, as for any mixture model, Eq. 5.6 becomes:

[2] Note that Wooldridge (2005) is more general in his approach and other distributions might be used. Here we assume $\Sigma_{\alpha_i} = \sigma^2_{\alpha_i} I$. In the current developments, the parameters of the normal distribution σ_{α_i} are location specific, but the i subscript is omitted to keep the notation simple.

$$P(y_{in1}, y_{in2}, ..., y_{int}|y_{in0}) = \int_{\alpha_n} \prod_{t=1}^{T} P(y_{int}|y_{in0}, y_{in(t-1)}, \alpha_n) f(\alpha_n|y_{in0}, \bar{x}_n) d\alpha_n. \quad (5.7)$$

Here, $P(y_{int}|y_{in0}, y_{in(t-1)}, \alpha_n)$ is a logit model. $f(\alpha_n|y_{in0}, \bar{x}_n)$ is normally distributed, following Eq. 5.4. Endogeneity is corrected.

Table 5.1 summarizes the three different models presented in Section 5.2.

Static model	Dynamic model without agent effect	Dynamic model with agent effect
$\rho = 0$	$\rho \neq 0$	$\rho \neq 0$
$a, b, c, \sigma_\alpha^2 = 0$	$a, b, c, \sigma_\alpha^2 = 0$	$a, b, c, \sigma_\alpha^2 \neq 0$

Table 5.1 – Description of static model, dynamic model without agent effect and dynamic model with panel effect graphically and as a function of Eq 5.5.

5.3 Pedestrian case study for EPFL catering locations

We present results for the three models presented in Section 5.2, summarized in Table 5.1, in the context of location choice on the EPFL campus. We focus on the choice of catering facilities during their opening hours. The choice set $\mathscr{C}$ contained 21 alternatives corresponding to the services available in 2012 (Fig. 5.1). We use the output of Ch. 3 as data for estimation and forecasting, with $L = 1$. Data and a dynamic model specification for Pythonbiogeme (Bierlaire; 2003; Bierlaire and Fetiarison; 2009) are available in Tinguely and Danalet (2015).

5.3.1 Model specification and estimation

The explanatory variables used for the location choice are (1) attributes varying with the alternatives: distance from the previous activity episode in the sequence, duration, cost, time of the day, opening hours, quality evaluation of the catering location, its capacity, its type of offer, and (2) characteristics describing the choice context, constant across alternatives: weather conditions, day of the year, socio-economic attributes. Descriptive statistics on the collected data are available in Appendix A.3.

Two lagged variables $y_{in(t-1)}$ are defined in the dynamic models, one for the morning and one for the lunch break. Thus, the dynamic Markov process is over individuals and periods of the

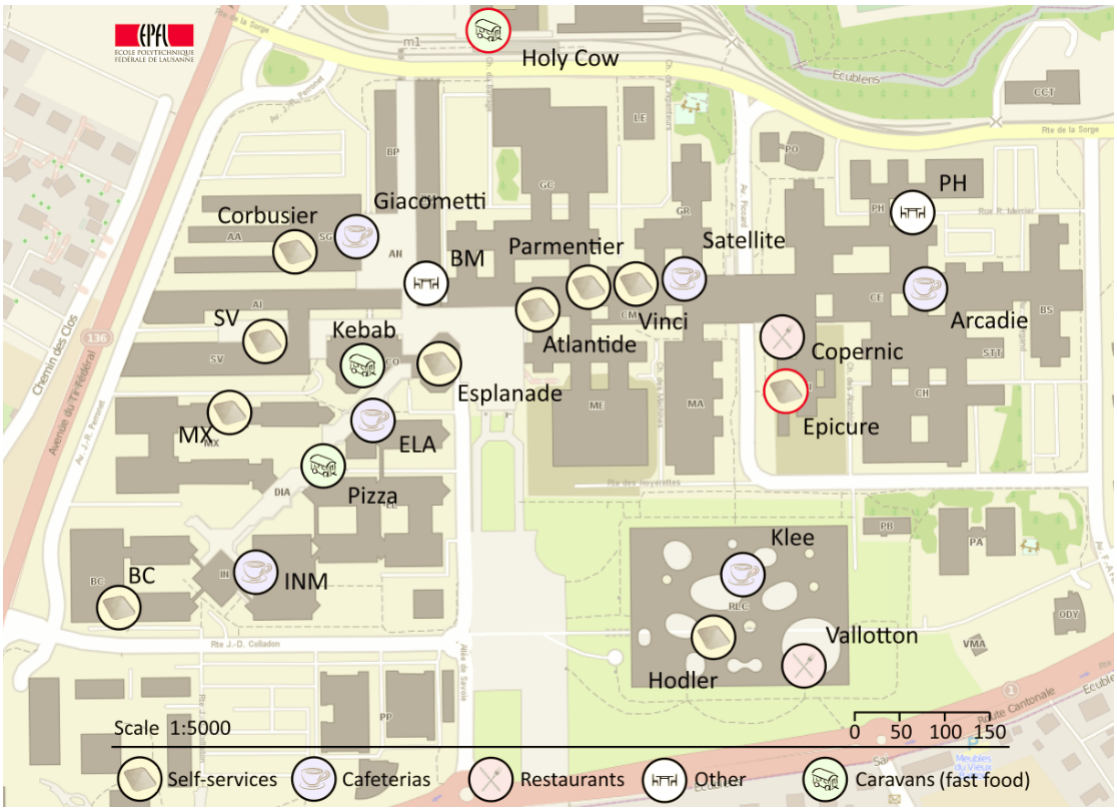


Figure 5.1 – Catering facilities on EPFL campus with different categories: self-services, cafeterias, restaurants, caravans (fast food) and others. The alternatives in red circles did not exist in 2012, when WiFi traces were collected. Image: Tinguely (2015).

day. Equations 5.2 and 5.3 become:

$$U_{int} = V_{int} + \rho_{\text{morning}} y^{\text{morning}}_{in(t-1)} + \rho_{\text{lunch}} y^{\text{lunch}}_{in(t-1)} + \varepsilon_{int} \tag{5.8}$$

$$U_{int} = V_{int} + \rho_{\text{morning}} y^{\text{morning}}_{in(t-1)} + \rho_{\text{lunch}} y^{\text{lunch}}_{in(t-1)} + \alpha^{\text{morning}}_{in} + \alpha^{\text{lunch}}_{in} + \varepsilon'_{int} \tag{5.9}$$

The specification of the agent effect distribution must be correct to get consistent estimates (Wooldridge; 2005). Therefore, we propose two different specifications for α_{in}. The first specification corresponds to $c = 0$ in Eq. 5.4. We assume the agent effect to depend only on the *first choice*:

$$\alpha_{in} = a + b y_{in0} + \xi_n. \tag{5.10}$$

The second specification for the agent effect includes the count y^{count}_{int} of previous choices of alternative i by individual n up to the event t of the current choice: $y^{\text{count}}_{int} = \sum_{t'=1}^{t-1} I(y_{int'})$. Note that in the definition of the count of previous choices, the first observation y_{in0} is not included

and the summation start at $t' = 1$. It allows to avoid biases (Rabe-Hesketh and Skrondal; 2013).

Eq. 5.4 becomes:

$$\alpha_{in} = a + by_{in0} + cy_{int}^{\text{count}} + \xi_n. \tag{5.11}$$

Since the lagged variable $y_{in(t-1)}$ is interacted with the period of the day, the count of previous choices is also specified for each period of the day.

Consequently, we estimate 4 models: the static model (utility defined in Eq. 5.1), a dynamic model without agent effect (utility defined in Eq. 5.2) and two dynamic models with agent effect: one with a first choice agent effect specification (Eq. 5.10) and one with a first choice and frequency specification (Eq. 5.11).

Static model	**Dynamic model without agent effect**	**Dynamic model with agent effect**	
		First choice	First choice and frequency
$\rho = 0$	$\rho \neq 0$	$\rho \neq 0$	$\rho \neq 0$
$a = 0$	$a = 0$	$a \neq 0$	$a \neq 0$
$b = 0$	$b = 0$	$b \neq 0$	$b \neq 0$
$c = 0$	$c = 0$	$c = 0$	$c \neq 0$
$\sigma_\alpha^2 = 0$	$\sigma_\alpha^2 = 0$	$\sigma_\alpha^2 \neq 0$	$\sigma_\alpha^2 \neq 0$

Table 5.2 – Description of static model, dynamic model without agent effect and two dynamic models with panel effect used in the case study as a function of Eq 5.5.

We use a linear specification for the different models, whose variables are described in Table 5.3. Table 5.4 describes the estimation results for the 4 models of Table 5.2. In this model, habits are assumed only for the morning and lunch break.

	Static model		**Dynamic model without agent effect**		**Dynamic model with agent effect**			
					First choice		First choice and frequency	
Parameters	Value	t-test	Value	t-test	Value	t-test	Value	t-test
$ASC_{\text{Le Klee}}$	−3.26	−5.52	−2.91	−4.82	−4.90	−3.75	−5.24	−3.83
$ASC_{\text{Cafétéria BC}}$	0.387	0.97*	0.481	1.12*	−1.09	−1.62*	−0.682	−0.88*
ASC_{BM}	0.450	1.29*	0.453	1.33*	−0.147	−0.24*	−0.320	−0.49*
$ASC_{\text{Cafétéria ELA}}$	−0.823	−2.42	−0.579	−1.59*	−1.08	−1.68*	−0.919	−1.67*
$ASC_{\text{Cafétéria INM}}$	−2.19	−3.97	−1.82	−3.13	−1.64	−1.52*	−1.81	−1.75*
$ASC_{\text{Cafétéria MX}}$	−0.461	−1.22*	−0.514	−1.23*	−1.89	−2.05	−1.78	−2.57
ASC_{PH}	1.28	3.48	1.11	2.99	0.298	0.62*	0.704	1.39*
$ASC_{\text{L'Arcadie}}$	−0.738	−2.08	−0.684	−1.85*	−1.98	−1.85*	−1.70	−1.81*
$ASC_{\text{L'Atlantide}}$	−0.143	−0.47*	−0.285	−0.88*	−1.23	−2.21	−0.731	−1.23*
$ASC_{\text{Le Copernic}}$	2.83	2.04	2.67	2.29	2.59	0.75*	1.88	1.25*
$ASC_{\text{Le Corbusier}}$	−0.278	−2.05	−0.259	−1.74*	−1.05	−2.52	−0.585	−2.28
$ASC_{\text{Le Giacometti}}$	0.323	1.12*	0.398	1.26*	0.760	1.47*	0.685	1.34*

continued ...

	Static model		Dynamic model without agent effect		Dynamic model with agent effect			
					First choice		First choice and frequency	
Parameters	Value	t-test	Value	t-test	Value	t-test	Value	t-test
$ASC_{\text{Le Parmentier}}$	−0.846	−3.22	−0.883	−3.14	−1.44	−3.63	−1.60	−3.61
$ASC_{\text{Le Vinci}}$	−4.11	−5.77	−3.81	−5.35	−8.24	−2.58	−4.97	−3.42
$ASC_{\text{L'Esplanade}}$	0.0	-	0.0	-	0.0	-	0.0	-
$ASC_{\text{L'Ornithorynque}}$	−0.631	−4.81	−0.641	−4.55	−1.26	−6.48	−1.24	−5.74
$ASC_{\text{Caravan Pizza}}$	−1.97	−3.40	−1.84	−3.23	−2.47	−2.89	−1.91	−2.79
$ASC_{\text{Caravan Kebab}}$	−2.73	−4.42	−2.51	−4.16	−3.12	−3.39	−2.64	−3.21
$ASC_{\text{Bar Satellite}}$	−1.60	−4.34	−1.42	−3.72	−2.27	−3.51	−2.65	−4.52
$ASC_{\text{Le Hodler}}$	0.995	2.07	0.954	2.10	2.40	3.33	2.76	3.86
$ASC_{\text{Table de Vallotton}}$	4.25	2.10	4.02	2.56	0.987	0.67*	1.34	0.80*
$\beta_{\text{dist, lunch, cafet}}$	−0.00689	−13.47	−0.00612	−11.64	−0.00406	−6.37	−0.00397	−6.71
$\beta_{\text{dist, lunch, rest}}$	−0.00138	−0.63*	−0.00127	−0.62*	−0.000498	−0.29*	0.00166	0.75*
$\beta_{\text{dist, lunch, self}}$	−0.00638	−15.45	−0.00543	−12.88	−0.00394	−8.91	−0.00400	−9.32
$\beta_{\text{dist, lunch, fast food}}$	−0.00953	−9.55	−0.00881	−9.06	−0.00672	−5.50	−0.00676	−5.31
$\beta_{\text{dist, lunch, other}}$	−0.00187	−2.20	−0.00100	−1.40*	0.000738	0.79*	0.000190	0.17*
$\beta_{\text{dist, morning}}$	−0.00557	−5.74	−0.00448	−4.59	−0.00405	−3.88	−0.00390	−3.60
$\beta_{\text{dist, after lunch}}$	−0.000453	−0.76*	−0.00107	−1.84*	−0.00101	−1.67*	−0.00107	−1.72*
$\beta_{\text{no dist}}$	−5.07	−12.70	−4.48	−11.79	−3.82	−6.98	−3.66	−7.66
$\beta_{\text{eval, cafet}}$	1.18	12.27	1.10	11.91	1.92	10.59	1.90	7.72
$\beta_{\text{eval, self}}$	1.21	9.25	1.09	8.45	2.12	8.28	2.02	6.55
$\beta_{\text{eval, fast food}}$	1.69	11.81	1.60	11.85	2.71	10.78	2.58	8.65
$\beta_{\text{cost, student}}$	−0.245	−3.50	−0.189	−3.01	−0.471	−4.00	−0.538	−4.47
$\beta_{\text{cost, employees}}$	−0.128	−2.26	−0.102	−1.97	−0.352	−3.20	−0.368	−3.64
β_{beer}	0.722	4.07	0.539	3.02	1.05	3.85	1.14	3.93
β_{dinner}	1.04	3.34	1.03	3.39	0.997	2.60	0.795	2.01
β_{capacity}	0.00680	2.62	0.00749	2.69	0.0104	2.71	0.0119	2.82
ρ_{morning}	0.0	-	3.06	17.48	0.591	1.09*	0.476	1.69*
b_{morning}	0.0	-	0.0	-	1.80	3.10	1.46	4.82
c_{morning}	0.0	-	0.0	-	0.0	-	0.450	2.76
ρ_{lunch}	0.0	-	1.78	15.45	0.644	4.36	0.355	1.95*
b_{lunch}	0.0	-	0.0	-	1.19	5.50	1.07	5.22
c_{lunch}	0.0	-	0.0	-	0.0	-	0.618	3.36
$\sigma_{\text{Klee, morning}}$	0.0	-	0.0	-	−2.55	−3.51	2.17	3.28
$\sigma_{\text{BC, morning}}$	0.0	-	0.0	-	1.76	5.91	1.61	3.90
$\sigma_{\text{BM, morning}}$	0.0	-	0.0	-	−0.578	−1.04*	0.195	0.51*
$\sigma_{\text{ELA, morning}}$	0.0	-	0.0	-	1.72	2.69	1.14	2.80
$\sigma_{\text{INM, morning}}$	0.0	-	0.0	-	−1.01	−2.31	0.725	0.87*
$\sigma_{\text{MX, morning}}$	0.0	-	0.0	-	−0.0850	−0.18*	1.17	1.91*
$\sigma_{\text{PH, morning}}$	0.0	-	0.0	-	0.246	0.79*	−0.352	−1.44*
$\sigma_{\text{Arcadie, morning}}$	0.0	-	0.0	-	1.41	1.77*	−0.726	−1.19*
$\sigma_{\text{Atlantide, morning}}$	0.0	-	0.0	-	−1.85	−6.59	1.21	4.39
$\sigma_{\text{Copernic, morning}}$	0.0	-	0.0	-	−0.922	−0.40*	0.378	0.86*
$\sigma_{\text{Corbusier, morning}}$	0.0	-	0.0	-	1.85	2.74	−1.74	−2.59
$\sigma_{\text{Giacometti, morning}}$	0.0	-	0.0	-	−0.00721	−0.02*	−0.147	−0.49*
$\sigma_{\text{Parmentier, morning}}$	0.0	-	0.0	-	0.967	1.69*	−1.43	−1.98
$\sigma_{\text{Vinci, morning}}$	0.0	-	0.0	-	−0.0815	−0.23*	0.396	0.74*
$\sigma_{\text{Esplanade, morning}}$	0.0	-	0.0	-	0.0	-	0.0	-
$\sigma_{\text{Ornithorynque, morning}}$	0.0	-	0.0	-	0.0261	0.05*	0.237	0.59*
$\sigma_{\text{Pizza, morning}}$	0.0	-	0.0	-	1.60	2.74	−0.932	−6.03
$\sigma_{\text{Kebab, morning}}$	0.0	-	0.0	-	1.82	5.98	−1.79	−5.59
$\sigma_{\text{Satellite, morning}}$	0.0	-	0.0	-	2.02	5.35	−2.32	−6.41
$\sigma_{\text{Hodler, morning}}$	0.0	-	0.0	-	1.71	2.42	0.290	0.41*
$\sigma_{\text{Vallotton, morning}}$	0.0	-	0.0	-	0.578	0.53*	0.292	0.75*
$\sigma_{\text{Klee, lunch}}$	0.0	-	0.0	-	−2.59	−5.44	2.71	7.08

continued ...

	Static model		Dynamic model without agent effect		Dynamic model with agent effect			
					First choice		First choice and frequency	
Parameters	Value	t-test	Value	t-test	Value	t-test	Value	t-test
$\sigma_{\text{BC, lunch}}$	0.0	-	0.0	-	2.06	6.11	−2.20	−7.48
$\sigma_{\text{BM, lunch}}$	0.0	-	0.0	-	2.33	3.52	−2.50	−4.18
$\sigma_{\text{ELA, lunch}}$	0.0	-	0.0	-	−1.05	−3.92	−0.789	−2.60
$\sigma_{\text{INM, lunch}}$	0.0	-	0.0	-	0.883	1.47*	−1.33	−2.83
$\sigma_{\text{MX, lunch}}$	0.0	-	0.0	-	−2.06	−6.27	1.66	9.55
$\sigma_{\text{PH, lunch}}$	0.0	-	0.0	-	2.63	7.13	−2.30	−3.39
$\sigma_{\text{Arcadie, lunch}}$	0.0	-	0.0	-	2.97	5.74	2.46	5.25
$\sigma_{\text{Atlantide, lunch}}$	0.0	-	0.0	-	1.85	5.44	−1.54	−6.82
$\sigma_{\text{Copernic, lunch}}$	0.0	-	0.0	-	5.78	2.92	6.06	4.32
$\sigma_{\text{Corbusier, lunch}}$	0.0	-	0.0	-	−1.27	−3.89	−0.855	−3.16
$\sigma_{\text{Giacometti, lunch}}$	0.0	-	0.0	-	−1.31	−6.13	1.24	6.35
$\sigma_{\text{Parmentier, lunch}}$	0.0	-	0.0	-	0.961	3.75	−1.19	−2.57
$\sigma_{\text{Vinci, lunch}}$	0.0	-	0.0	-	3.56	1.91*	−1.37	−1.36
$\sigma_{\text{Esplanade, lunch}}$	0.0	-	0.0	-	0.0	-	0.0	-
$\sigma_{\text{Ornithorynque, lunch}}$	0.0	-	0.0	-	0.128	0.49*	−0.258	−1.23*
$\sigma_{\text{Pizza, lunch}}$	0.0	-	0.0	-	−1.24	−5.42	1.29	5.15
$\sigma_{\text{Kebab, lunch}}$	0.0	-	0.0	-	0.677	3.00	−1.11	−4.48
$\sigma_{\text{Satellite, lunch}}$	0.0	-	0.0	-	0.776	5.26	−1.20	−4.13
$\sigma_{\text{Hodler, lunch}}$	0.0	-	0.0	-	1.05	3.51	−0.910	−1.91*
$\sigma_{\text{Vallotton, lunch}}$	0.0	-	0.0	-	10.7	5.52	−10.8	−7.20
Nb of observations	1868							
$\mathscr{L}(0)$	−5037.914							
Nb estim. param.	36		38		80		82	
$\mathscr{L}(\hat{\beta})$	−3446.109		−3092.106		−2631.929		−2623.843	
Adjusted rho square $\bar{\rho}^2$	0.309		0.379		0.462		0.480	
Likelihood ratio test	354.003 (> 5.99)		920.354 (> 58.12)		16.172 (> 5.99)			

Table 5.4 – Summary of estimation results for the 4 models of Table 5.2. 1868 observations are used for estimation. Parameters without stars are significantly different from zero with a 95 % confidence level. A likelihood ratio test is performed between the static model and the dynamic model without agent effect, between the dynamic model without agent effect and the dynamic model with agent effect (first choice specification), and between the dynamic model with agent effect (first choice specification) and the dynamic model with agent effect (first choice and frequency). The numbers in parenthesis for the likelihood ratio tests are the percentiles of the χ^2 distribution.

Lagged variables ρ_{lunch} and ρ_{morning} have positive signs, showing habits and repeated choices. Their value decreases when the dynamic model includes the agent effect (as compared to the model when it is considered exogenous). It has been reported in Monte Carlo simulations that ρ is overestimated in dynamic models without agent effect as compared to dynamic models with agent effect (Akay; 2012). This is due to the double nature of the lagged variable ρ: the previous choice impacts the current choice because the past experience modifies the current preferences and because the past and current choices both depend on the same time-persistent unobserved parameters. These two factors are called *true state dependence* and *spurious state dependence*, respectively, by Heckman (1978, 1981) (see also Hsaio; 2003, Section 7.5.4). The agent effect, and in particular the *first choice and frequency* version of it, is

Parameters	Variables	Description of the variable
ASC_i	1_i	Alternative specific constant for catering location i
$\beta_{\text{dist, cat},ToD}$	$\text{dist}_{\text{cat},ToD}$	Distance from the previous activity episode
$\beta_{\text{no dist}}$	$1_{\text{dist } NA}$	Variable for missing data about distance
β_{eval}	eval_i	Evaluation of catering location from survey data (grade between 1 and 6)
$\beta_{\text{cost, student}}$	$\text{cost}_{\text{students}}$	Cost of the cheapest meal for students
$\beta_{\text{cost, employees}}$	$\text{cost}_{\text{employees}}$	Cost of the cheapest meal for employees
β_{beer}	1_{beer}	Availability of beer after 14:00
β_{dinner}	1_{dinner}	Availability of dinner
β_{capacity}	$\text{capacity}_{\text{outdoor}}$	Number of seats in the catering location
ρ_{ToD}	$y_{it(t-1),ToD}$	Indicator variable with value 1 if the previous catering location in the same time of day (ToD) is the same as the current catering location
c_{ToD}	y_{int}^{count}	Variable counting the frequency of visit to catering location i in the same time of day (ToD)
$\sigma_{i,ToD}$	1_{ToD}	Variance of ξ for each time of day (ToD) and each catering location i

Table 5.3 – Description of the variables in the catering location choice model. Some variables are interacted with the category of catering location (cat), i.e. the categories of restaurant presented in Fig. 5.1, or are interacted with time of day (ToD), divided in morning (until 11:29), lunch break (11:30-13:59), afternoon (14:00-17:59), dinner (18:00-19:59) and night (from 20:00).

absorbing the time-persistent unobserved preferences.

The parameters have expected signs. Indoor capacity (number of seats) has a positive impact on the choice of visiting a catering location. Distance from the previous activity episode has a negative impact on the propensity to visit a catering location. This effect is strong in the morning and during lunch time for cafeterias, while it is not significant in the afternoon and during lunch time for restaurants (there are not many restaurants on campus, and consequently longer distances to walk). The cost parameters have a negative sign and their magnitude is larger for student than for employees. This is explained by the fact that employees have salaries and thus a higher purchasing power and a lower sensitivity to price. Annual evaluations by students (as a proxy for average quality), offering meals for dinner and beers after 14:00 all have a positive impact on the choice of catering locations.

The dynamic model without agent effect, the dynamic model with agent effect (*first choice* correction) and the dynamic model with agent effect (*first choice and frequency* correction) are unrestricted versions of the previous, simpler model in Table 5.2 (i.e., static model, dynamic

model without agent effect, dynamic mode with agent effect (*first choice* correction), resp.). Table 5.4 shows the results of three likelihood ratio tests. In all cases, we can reject the null hypothesis at a 95 % confidence level and the unrestricted model is preferred to the restricted one.

5.3.2 Validation

Cross-validation has been performed, partitioning the data in an *estimation dataset* containing past observations $i_1, i_2, ..., i_{T_n-1}$ of individuals n and a *validation dataset* with their most recent choice i_{T_n}. Models presented in Section 5.3.1 are applied to observations in the morning and during lunch break, in order to test the dynamics. The estimation dataset contains 1512 observations. The model is then applied to the validation dataset (containing 144 observations), using the parameter estimates from the previous step. Aggregate average number of visits across individuals' most recent choices from observations and from the model output are compared in Table 5.5.

In order to compare the performance of the different models over all catering locations in Table 5.5, we compute the sum of the squares of the errors: $S_m = \sum_i (O_i - E_{i,m})^2$, where O_i is the observed average number of visits for location i and $E_{i,m}$ is the expected average number of visits based on the choice probabilities for location i assuming model m.

Observed and predicted average number of visits show similar tendencies, even for the static model, meaning that the specification of the model is generally good. The model minimizing the sum of the squares of the errors is the dynamic model with agent effect using the first choice and the frequency. It is also the model that fits the data the best (Table 5.4). It is an evidence that Wooldridge's approach is valid, and it performs better when the specification of the agent effect distribution includes the frequency of visits.

5.3.3 Elasticity to price

Aggregate direct elasticity of cost denotes the percent change in the number of visits for each catering location with respect to a change of 1 % in the cost of a meal. Aggregate direct elasticities of cost are presented for each restaurant, for students and employees, in Table A.4 in Appendix A.4. Figure 5.2 summarizes the distribution of aggregate direct elasticities of cost as box-plots for each model, across students and employees.

Demand for catering locations for students is more elastic to a change in the cost of a meal as compared to employees. This is explained by the higher purchasing power of employees. With the static model and the dynamic model without agent effect, employees mostly show a inelastic demand (< 1 in absolute value) and students show an elastic demand (> 1 in absolute value). With the dynamic models with agent effect, using the first choice and the frequency of choices, the absolute values of elasticities increase and employees have an elastic demand with respect to the cost of a meal. Generally, models ignoring the dynamics are less sensitive to

	Observed		Predicted							
			Static model		Dynamic model without agent effect		Dynamic model with agent effect			
							First choice		First choice and frequency	
Catering locations	Nb	%	Nb	%	Nb	%	Nb	%	Nb	%
Cafet. Le Klee	0	0.0	0.4	0.3	0.3	0.2	0.4	0.3	0.3	0.2
Cafet. ELA	14	9.7	7.6	5.3	6.9	4.8	8.0	5.5	8.0	5.6
Cafet. INM	1	0.7	1.2	0.9	1.1	0.8	2.2	1.5	2.1	1.4
Cafet. MX	6	4.2	6.3	4.4	6.4	4.4	5.3	3.7	5.8	4.0
Cafet. L'Arcadie	6	4.2	1.4	1.0	2.4	1.7	1.5	1.1	1.7	1.2
Cafet. Le Giacometti	13	9.0	12.0	8.3	11.8	8.2	12.8	8.9	12.2	8.5
Cafet. Satellite	5	3.5	7.2	5.0	7.6	5.3	7.8	5.4	7.5	5.2
Self BC	15	10.4	9.7	6.7	9.5	6.6	10.8	7.5	10.8	7.5
Self L'Atlantide	7	4.9	10.8	7.5	10.6	7.4	8.2	5.7	8.1	5.6
Self Le Corbusier	4	2.8	12.6	8.7	10.6	7.4	9.4	6.5	10.8	7.5
Self Le Parmentier	8	5.6	13.1	9.1	12.9	9.0	13.1	9.1	13.2	9.1
Self Le Vinci	1	0.7	0.2	0.1	0.2	0.1	0.2	0.1	0.2	0.1
Self L'Esplanade	23	16.0	26.1	18.2	25.9	18.0	24.2	16.8	24.4	17.0
Self L'Ornithorynque	15	10.4	15.0	10.4	16.4	11.4	15.6	10.8	15.7	10.9
Self Le Hodler	6	4.2	5.2	3.6	6.1	4.3	5.7	4.0	6.2	4.3
Rest. Le Copernic	1	0.7	1.0	0.7	1.4	1.0	3.4	2.4	3.3	2.3
Rest. Table de Vallotton	1	0.7	1.3	0.9	1.1	0.8	0.6	0.4	0.5	0.3
Caravan Pizza	6	4.2	4.2	2.9	4.5	3.1	4.5	3.1	4.8	3.4
Caravan Kebab	5	3.5	3.6	2.5	3.7	2.6	3.5	2.4	3.8	2.6
Other BM	1	0.7	1.8	1.2	1.2	0.8	1.6	1.1	1.3	0.9
Other PH	6	4.2	3.2	2.2	3.6	2.5	3.3	2.3	3.3	2.3
S_m			232.95		204.01		184.16		173.85	

Table 5.5 – Aggregate average number of visits of the observations and of the different models, from the 144 most recent observations for each individual in the morning and during lunch break. For the observations and for each model, the number of visitors ("Nb") and the proportion of visitors ("%") are presented for each catering location. "Rest." stands for restaurant, "Self" for self-service, "Cafet." for cafeteria.

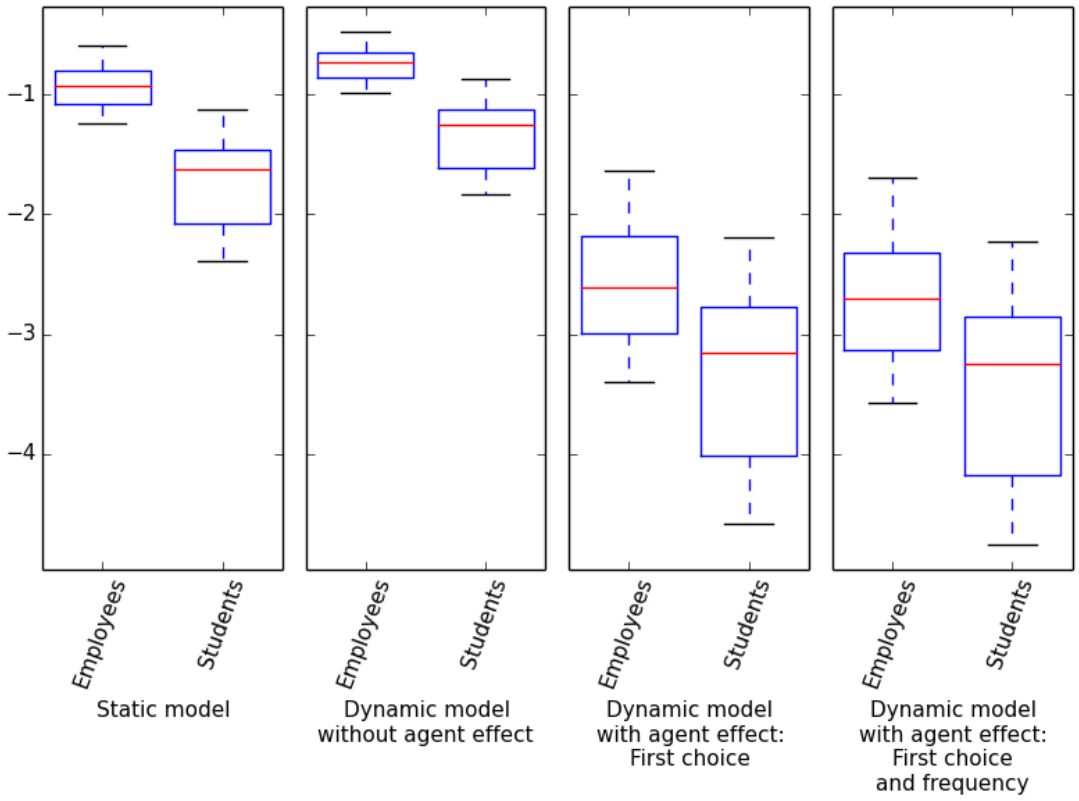


Figure 5.2 – Distribution of aggregate direct elasticities of cost for different models, for students and employees.

cost. A possible analogy is the presence of unobserved variables, such as quality of the service or of the meal (Train; 2003, ch. 13). Decision makers prefer cheap meals, but also like quality meals. When endogeneity is not corrected for, β_{cost} absorbs both effects and its absolute value is attenuated. When endogeneity is accounted for, β_{cost} is more negative, including only the taste for cheap meals. Here, in the static model and the model without agent effect, β_{cost} absorbs a taste for cheap meals and other unobserved factors positively correlated with cost, such as a warm atmosphere or any attribute of quality for a meal. In the models with agent effect, unobserved factors are absorbed by the agent effect.

5.3.4 Forecasting visits when opening a new catering location

Data used for estimation have been collected in 2012. We forecast the average number of visits for 2013 after the opening of a new self-service.

In this scenario, habits regarding the new catering location are not considered. The new alternative is not part of people's habits in the model: the previous catering location and the frequency of visits are null ($y_{in(t-1)} = 0$ and $y_{int}^{\text{count}} = 0$ when i is the new catering location).

A new self-service, *L'Epicure*, opened in October 2013. The four models of Table 5.2 are applied to this new choice set. The parameters for the new self-service are the same as *Le Giacometti*, since it is the most similar existing catering location on campus and no stated preference is available.

The error term of the new alternative and the error term of the most similar existing alternative might be correlated. Indeed, if the new catering location does not share any unobserved attribute with the most similar catering location, a logit specification is valid. On the contrary, if unobserved attributes are shared, the two locations should be included in a nest and a nested logit specification is used for forecasting. Since we don't know the value of the nest parameter θ, an interval of values is used from 1 (i.e., logit model and independent error terms) to $+\infty$ (i.e., perfectly correlated error terms) when applying the model to forecast average number of visits. Results are presented in Fig. 5.3.

When using a static model, the predicted average frequency of visits varies between 0.7 % and 2.0 % for the full day. When correcting for endogeneity and using frequency of visits in the specification of the agent effect, the predicted average frequency of visits varies between 0.4 % and 1.1 %. It shows that correcting for endogeneity when using panel data has a significant impact when predicting the destination choices of people. The effect of the unknown level of correlation between the new catering location and its most similar alternative also seems lower when using the dynamic models with agent effect.

According to point-of-sale data collected from October 21 to 23, 2013, the frequency of financial transactions in the new self-service is 1.5 %, that has a level of magnitude consistent with the values predicted by the model.

Results presented here are only valid in the short term, since the model has been applied only once. For forecasting in the long term, accounting for the habits and routines, the model should be applied several times so that habits for the new location are establishing as an output of the model.

5.4 Conclusion

In this chapter, we model location choice conditional on the choice of activity type in activity episodes from WiFi traces. WiFi traces provide panel data. We estimate dynamic models, including lagged variables. They express the habits that could appear in repeated choices.

Including lagged variables in a discrete choice model generates endogeneity. The error term and the explanatory variable representing the previous choice are serially correlated. The so-called initial conditions problem is solved using a control function proposed by Wooldridge (2005). The error term is decomposed in an agent effect and an independent error term. The conditional distribution of the agent effect, knowing the first choice, is approximated.

The approach of Section 5.2 has been applied to a case study on a campus (Section 5.3), based

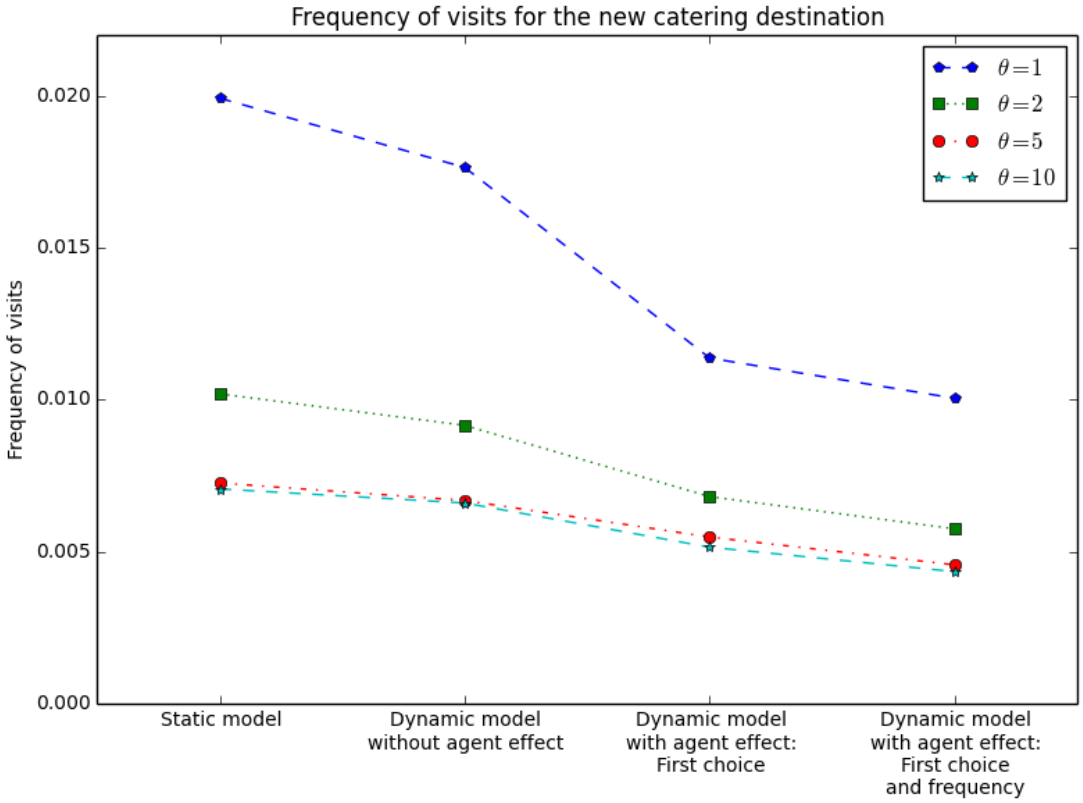


Figure 5.3 – Average frequency of visits for the new self-service for the different models, as a function of θ.

on actual WiFi traces, preprocessed as in Chapter 3. Campus members tend to visit catering locations that are closer, with large capacities, that offer beers and serve meals for dinner. Students are more sensitive to cost than employees. The previous choice significantly impacts the choice of the current catering location in the morning and in the lunch break in a dynamic model without agent effect, without the correction for endogeneity. When controlling for true state dependence and spurious state dependence, time-persistent unobserved effects are detected and the previous choice becomes not significant anymore. A likelihood ratio test has been performed between the different models, and the null hypothesis of a restricted model can always be rejected when comparing two consecutive models in terms of complexity: a dynamic model without agent effect (without correction for endogeneity) is preferred to a static model, a dynamic model with agent effect is preferred to the dynamic model without agent effect, and an agent effect including the first choice and the frequency in its specification is preferred to an agent effect specified with the first choice only.

Models are validated in Section 5.3.2 and the models seem correctly specified, reproducing the observations of the validation dataset. In terms of predictive power, dynamic models outperform static models, and the agent effect including the first choice and the frequency of visits performs the best.

Elasticity to the cost of a menu and forecasting in the case of the opening of new catering locations are presented in Section 5.3.3 and 5.3.4, respectively. Elasticity to the cost of a menu increases with dynamic models with agent effect. In the scenario of the opening of a new catering location, predicted average number of visits correspond to point-of-sale data of the first week of opening.

This model can be applied in pedestrian facilities to estimate demand for specific locations. Wooldridge's approach is easy to implement for discrete choice models with many alternatives and improves the estimation and predictive power of the model. Our model specification could be extended towards more complex discrete choice models (e.g., a nested logit where categories of catering locations would be the nests in our case study). Collection of more socioeconomic data would also improve the specification and prove useful for marketing purposes. On campuses, in transportation hubs or music festivals, information on congestion at location (i.e., queues for a service) is likely to be significant in explaining people's behavior. Endogeneity in the model due to congestion could also be corrected, using the occupation rates for each location as measures of queues and congestion at these locations. Some endogeneity could also be related to group effects, when a group chooses a location together instead of each individual independently (Louviere et al.; 2005, Section 2). This could be corrected using proximity as a measure of social networks. Finally, space syntax has been used in recent research and could help in formalizing intuitions such as "visibility" in public spaces.

6 Conclusion

In this concluding chapter, we first present empirical findings and discuss theoretical and policy implications in Section 6.1, then we outline future research directions and describe some limitations in Section 6.2.

6.1 Empirical findings, theoretical and policy implications

This thesis proposes an activity-based model of pedestrian demand using WiFi traces. Stop and activity at the stop are extracted by merging the WiFi localization data with other data sources. Then, the choice of activity type, duration and time of day is simultaneously modeled. Finally, a dynamic model of location choice among all alternatives for a given activity type includes panel effect. The different steps to reach a complete activity-based pedestrian demand model are presented in Fig. 1.1, page 5.

The empirical findings of this dissertation show that it is possible to detect and model pedestrian activity-episode sequences from WiFi traces. In this section, we summarize these empirical findings. We evaluate the impact of our contributions on existing models, for pedestrian behavior and more generally for activity-based modeling. We also provide policy implications of the empirical findings for pedestrian facilities.

6.1.1 Detecting stops and activities performed at these stops

In Ch. 3, the different activity locations visited by a device are detected from imprecise localization data with dense points of interest. A Bayesian approach integrates information from time constraints, such as schedules, and from aggregate measures of occupation, such as point-of-sale data. Our contributions include defining activity performed at the stop from map data, dealing with overlapping antenna coverage and improving localization using other data sources.

Defining activity performed at the stop from map data

When using sensors data in pedestrian context, the activity location is often associated with the sensor. Delafontaine et al. (2012) associate a letter in a sequence with each Bluetooth sensor and Versichele et al. (2012) locate the Bluetooth sensors in strategic locations and consider that the device is located at the sensor location. In the WiFi literature, the goal is specifically to detect movements between access points (APs) (see Section 2.2.3). Contrarily to the rest of the WiFi literature, Yoon et al. (2006) aggregate APs at the building level and consider the building as the stop. At the urban scale, using GPS or GSM data, land-use data are also used to detect the activity purpose (Miller; 2014), but face the issue of mixed land-use compared to the precision of the measurement (Rieser-Schüssler; 2012).

In Ch. 3 of this dissertation, we assume that a sensor can cover several possible activity locations. We collect this information from the map. It is then associated with the measurements using *a priori* knowledge of their occupation (*potential attractivity measure*) and distance to the measurement (measurement equation). We believe this approach has more behavioral meaning. Our approach is formally well defined and does not need participant input or manual checks. It also does not need to locate the sensors at specific locations but allows to use already installed sensors. It removes the bias of locating the sensors in strategic locations defined by the analyst. Our approach, mixing localization data with map data and time constraints, could also be used at the urban scale.

Dealing with overlapping antenna coverage

When considering the sensor as the activity location, the device can rapidly change to which sensor it connects. This does not correspond to real movements but to technical changes in the signal strength. In the pedestrian context, Delafontaine et al. (2012) propose to decompose the time in units of 3 min and consider the sensor with the maximum share of this time unit as being the activity location. In WiFi literature, this is called the *ping pong* effect (see Section 2.2.3) and similar strategies have been used (moving average weighted by time spent at destination in Yoon et al.; 2006).

Our methodology avoids the pingpong effect by associating points of interest to localization measurements. We assume the device to be in an area (the *domain of data relevance*) surrounding the localization measurement. If localization measurements move (due to ping pong effect or measurement errors) while the device is in fact static, the true activity location is contained in both domains of data relevance and does not change.

Improving localization using other data sources

Mobile phone tracking data from already deployed WiFi networks are cost effective and do not require the installation of access points (see Section 2.1.1). However, they suffer from low precision. Other pedestrian counting data are used: manual counts, mechanical counts,

video counts, smart card data, etc. (see Section 2.1). These two types of data are usually studied separately. Kholod et al. (2010) use RFID data to track people in a grocery store and compare these results with point-of-sale data to study purchase behavior. In land-use planning literature, accessibility measures are common and represent the same idea as count data: a measure of the "size of the activity" (Hansen; 1959). Miller (2010) more recently adds constraints in a new definition of accessibility. To our knowledge, such data have not been associated with localization traces.

We define a *potential attractivity measure* for pedestrian facilities merging count data (e.g., point-of-sale data) and time constraints (e.g., schedules or opening hours). The potential attractivity measure is used as a prior in a Bayesian approach. An empirical study on the EPFL campus shows that this Bayesian approach makes up for the localization weakness and the dense pedestrian map. This methodology is robust for low density measurements. It is still theoretically valid in other contexts, such as for merging GSM data with land-use data at the urban scale.

6.1.2 Modeling the full activity pattern

In Ch. 4, we model the choice of a full activity pattern, represented as a path in a network. Our contributions include simultaneously modeling time of day preferences, duration and number of episodes, and overall pattern structure, managing the large choice set of activity paths and getting rid of home-based, tour-based structure and *a priori* assumptions.

Simultaneously modeling time of day preferences, duration and number of episodes, and overall pattern structure

For pedestrian behavior, Borgers and Timmermans (1986b) sequentially model the choice of a destination in time. Their approach does not include the preference an individual might have for a certain number of destinations in the sequence, preferences for certain durations, schedule constraints, or preferences for a certain order of activities. One possible approach to model the order of activities consist in comparing all possible combinations of activities. Hoogendoorn and Bovy (2004) propose such an approach and define an activity scheduling cost to impose mandatory activities, schedule constraints and the order of activities. Duration and number of activities are not considered. Section 2.3.3 reviews the large variety of models for activity-based model for a day, at the urban scale. Ettema et al. (2007) simultaneously model time-of-day preferences and duration, without considering the order in which activity episodes are performed. Habib (2011) models the scheduling of activities but not their planning, such as priorities of activities. Pinjari and Bhat (2010) model duration of activities but not their order. The activity-schedule approach (among others Shiftan and Ben-Akiva; 2011; Abou-Zeid and Ben-Akiva; 2012; Gupta and Vovsha; 2013) models time-of-day preferences, duration, priorities of activity and their order in different submodels.

We simultaneously model the choices of activity type, duration and time-of-day. We associate a utility with an activity path in an activity network, representing all possible combinations of time units and activity types. It allows to specify elements of the utility related to the full pattern, typically preferences for a primary activity, schedule delay effects, or patterns (e.g., a ticket purchase-platform pattern). The case study presents preferences for time of day, duration and number of episodes. It can include schedule delay effect, order constraints, primary activity and notions of priorities, such as mandatory activities. All these elements are estimated simultaneously.

Managing the large choice set of activity paths

Considering different orders for the activity episodes quickly increases the number of possible alternatives to be chosen. In a pedestrian context, Hoogendoorn and Bovy (2004) present an example with 2 consecutive activities. Liu, Usher and Strawderman (2014) define 3 time intervals. At the urban scale, Ettema et al. (2007) divide the day in three periods and assume work as the main central activity. In Pinjari and Bhat (2010) and Habib (2011), the order of activity episodes is not explicitly modeled. The activity-schedule approach sequentially models the order of activities and the detailed tours (see Section 2.3.3).

Modeling activity paths includes the order and the duration in a unique model. The large choice set is managed through recent developments in route choice modeling and in importance sampling strategies. Metropolis-Hastings sampling of paths (Flötteröd and Bierlaire; 2013) associated with strategic sampling (Lemp and Kockelman; 2012) generates a choice set of activity paths for importance sampling. The case study shows that our approach allows to include more parameters in the model than simple random sampling.

Getting rid of home-based, tour-based structure and *a priori* assumptions

Existing models assume postulated rules. The activity-schedule approach is structured on home and tours from home, with models applied sequentially according to priorities of activity types. Tours are motivated by one primary trip purpose, i.e., a main activity type per tour (Shiftan; 1998; Limanond et al.; 2005). Doherty and Mohammadian (2011) show that the decomposition between mandatory (e.g., work or school) and discretionary activity is not supported by data; they estimate an ordered response logit for activity sequences and show that duration explains activity planning more than activity type.

By managing the large choice set of activity paths, we do not assume tours nor priorities between activities. The utility of an activity path can include or not a home activity, a tour structure or a primary activity. Mandatory and discretionary can be characteristics of activity types but do not need to be defined and can be tested for significance.

6.1.3 Modeling location choice

Chapter 5 models location choice conditional on the activity-episode sequence using the panel data from WiFi traces. The methodology is applied to the choice of catering location on the EPFL campus. Our contributions consist in including lagged variables and correcting for serial correlation.

Including lagged variables

In pedestrian context, Ton (2014) and Kalakou et al. (2014) propose location choice models. They do not include the previous choice as an explanatory variable of the current choice. At a larger scale, Sivakumar and Bhat (2007) include a lagged variable in the choice of location. It represents a learning effect, due to update in preferences, delayed effects, variety seeking, habit persistence or loyalty behavior. In tourism literature, Grigolon et al. (2014) include the previous vacation length choice in the choice of the current vacation length. They compare a logit, a mixed logit and a dynamic mixed logit and show that the dynamic mixed logit is the best in estimation and forecasting.

We include lagged variables in the modeling of pedestrian activity location. Including preceding choice is possible due to the availability of activity-episode sequences for several days from WiFi traces. In terms of predictive power, dynamic models outperform static models

Correcting for serial correlation

Only few dynamic models of location choice exist in the literature, and none of them to our knowledge correct for serial correlation. Sivakumar and Bhat (2007) and Grigolon et al. (2014) do not consider this issue. In their dynamic mixed logit, by assuming that the error term is independent of the variables (i.e., exogenous), and in particular independent of the lagged variable, they assume that unobserved attributes do not persist over time for a given individual.

Including lagged variables in a discrete choice model generates endogeneity. The error term and the explanatory variable representing the previous choice are serially correlated in dynamic models. We apply Wooldridge (2005) method to deal with this endogeneity problem. The model correcting for endogeneity outperforms models not correcting for it. Predicted market shares of the new catering location, not open when WiFi traces were collected, correspond to point-of-sale data of the first week of opening.

6.1.4 Policy implication

Results show that we can detect and model activity episodes using existing tracking data. With a growing number of visitors, such methodologies are necessary for pedestrian infrastructures for

- testing scenarios when modifying or building pedestrian infrastructures, designing public transit timetable or optimizing the spatial distribution of facilities;
- counting visitors, their duration of stay and their daily and hourly patterns of visits;
- marketing and evaluating the economic viability of facilities; and
- managing congestion.

In particular, multimodal transport hubs, such as airports or train stations, face increasing demand and increasing commercial and leisure activities. Investments to modify them are considerable and potential conflicts between traveling and shopping activities must be avoided.

Network traces such as WiFi signatures are cheap to collect and easily cover the full infrastructure for long periods. Other sources of data are also available, such as point-of-sale data, train occupation or any aggregate measures of occupation.

Our methodology to detect activity-episode sequences is useful to merge the different data sources generated by pedestrian facilities. It allows to give behavioral sense to WiFi traces and observe activity-episode sequences in facilities. The activity path approach we propose allows to model the preference for certain activity types at certain times of day, the duration of activity episodes and their number. It provides information on the patterns performed by visitors, e.g., in train stations. Moreover, by including schedule delay effect in the model, it predicts activity pattern shifts related to a change in schedules, e.g., the impact of new timetables in train stations. The activity location choice model can predict the number of visitors in a future new location, as showed in the case study. It can also describe the preferences driving people's choice of location, such as the impact of price, distance, or general quality levels.

6.2 Future work and limitations

The methodology proposed for detection of activity-episode sequences in Ch. 3 should be applied in another setting than a campus, typically in a train station, where congestion and localization of facilities are important issues for daily management (e.g., an extension of Hänseler et al.; 2015 from two pedestrian underpasses to the full train station, using WiFi traces). It could also be applied to a larger scale, merging GSM data and land-use data (e.g., Bengtsson et al.; 2011 and Elwood et al.; 2012 in Haiti).

In Ch. 4, choice set generation and forecasting for activity path use a Metropolis-Hastings algorithm for the sampling of paths. The algorithm was originally developed for route choice and implicitly assume that the main attribute of the choice is distance. Activity path choice cannot be mostly explained by node-additive variables, i.e., time-of-day preferences. The dense activity network with constraints about the number of activity episodes requires to generate a large number of activity paths and discard a lot of similar paths to be able to assume

independence. The algorithm therefore needs time to generate choice sets. A useful future work would be to develop a more efficient Metropolis-Hastings algorithm for sampling of activity paths. It would drag an activity episode instead of a node. It could increase the time efficiency of the choice set generation process.

A strong assumption of the activity path approach in Ch. 4 is the usage of the logit model. The logit model is restricted by the Independent from Irrelevant Alternative (IIA) property. It might not be appropriate in the activity path approach since activity paths share unobserved attributes due to overlaps. Overlaps correspond to performing the same activity type at the same time and might be correlated. The well-known Path Size correction term in route choice modeling cannot be applied here. The choice of an activity path can be seen as an aggregation of alternatives (about aggregation of alternatives, see Ben-Akiva and Lerman; 1985, ch.9). The elemental alternatives are the activity paths, and the aggregate alternatives are the nodes in the activity network. In the route choice context, the Path Size formulation defines the size of an aggregate alternative (here, a link) as the number of paths using the link (Frejinger and Bierlaire; 2007). The motivation for this definition of size is related to the assumption that a physical overlap "measures" shared unobserved attributes. In the activity path context, the number of paths using a node is constant, K^{T-1}, due to the structure of the activity network, and so the Path Size formulation from route choice leads to a constant correction term for each alternative and consequently no correction. Fundamentally, this result comes from the symmetry of the activity network and the hypothesis that duration of the same activity type at the same time of day is a measure of similarity, i.e., one replaces physical shared length (overlap) from the route choice context by shared time length (duration) for a given activity type at a given time of day. Deterministic corrections of the utility for correlation, similar to the Path Size, should be developed to replace the Path Size logit in the context of activity path choice, e.g., based on the assumption that similarity for activity patterns is measured through shared primary activity purpose and pattern (Bowman; 1998). Stochastic corrections such as cross-nested logit could also be estimated (see Lai and Bierlaire; 2015).

In this dissertation, we decompose the modeling of activity-episode sequences in an activity path choice and a conditional activity location choice. Activity episodes are grouped into activity types in the activity network. Future work could model destination path choice, merging these two models. A destination path choice would define a destination network, i.e. the simultaneous choice of a series of activity episodes described by their time unit and activity location. The choice set would be larger than in the activity path choice.

The modeling results in Ch. 4 and 5 are based on the output of Ch. 3 considering only the most likely activity-episode sequence, i.e. with $L = 1$. It would be interesting to increase the number L of activity-episode sequences detected for one individual and one day and evaluate the impact on estimation of the activity path and activity location choices.

In future research and application, this thesis will generally help detect stop and activities performed at these stops merging localization data and land-use data. The methodologies

presented in this thesis also provide a framework for modeling the full activity pattern and for modeling location choice with longitudinal data.

A Appendix

A.1 Derivation of the distribution of t^-_{i+1}

Let's first assume $\hat{t}_i \leqslant \hat{t}_{i+1} - tt_{x_i,x_{i+1}}$ and $t^+_i + tt_{x_i,x_{i+1}} \leqslant \hat{t}_{i+1}$ to get rid of the maximum and the minimum in the bounds of the intervals, and thus simplify the notation (if these two conditions are not met, the two random variables t^+_i and t^-_{i+1} are fixed and the derivation is obvious).

The end time t^+_i is uniformly distributed, $t^+_i \sim U(\hat{t}_i, \hat{t}_{i+1} - tt_{x_i,x_{i+1}})$, with density function $f_{t^+_i}(x) = \frac{1}{\hat{t}_{i+1} - tt_{x_i,x_{i+1}} - \hat{t}_i}$ for $x \in [\hat{t}_i, \hat{t}_{i+1} - tt_{x_i,x_{i+1}}]$ and 0 otherwise. The start time t^-_{i+1} is uniformly distributed between $t^+_i + tt_{x_i,x_{i+1}}$ and $\hat{t}_{i+1}$. Its density for a given value of t^+_i is $f_{t^-_{i+1}|t^+_i=x}(y) = \frac{1}{\hat{t}_{i+1} - t^+_i - tt_{x_i,x_{i+1}}}$ for $y \in [x + tt_{x_i,x_{i+1}}, \hat{t}_{i+1}]$ and 0 otherwise. Now, the density of t^-_{i+1} is:

$$
\begin{aligned}
f_{t^-_{i+1}}(y) &= \int_{x=\hat{t}_i}^{\hat{t}_{i+1} - tt_{x_i,x_{i+1}}} f_{t^-_{i+1}|t^+_i=x}(y) \cdot f_{t^+_i}(x)\,dx && \text{(A.1)} \\
&= \int_{x=\hat{t}_i}^{y - tt_{x_i,x_{i+1}}} f_{t^-_{i+1}|t^+_i=x}(y) \cdot f_{t^+_i}(x)\,dx && \text{(A.2)} \\
&= \int_{x=\hat{t}_i}^{y - tt_{x_i,x_{i+1}}} \frac{1}{\hat{t}_{i+1} - x - tt_{x_i,x_{i+1}}} \cdot \frac{1}{\hat{t}_{i+1} - tt_{x_i,x_{i+1}} - \hat{t}_i}\,dx && \text{(A.3)} \\
&= \frac{1}{\hat{t}_{i+1} - tt_{x_i,x_{i+1}} - \hat{t}_i} \ln\left(\frac{\hat{t}_{i+1} - tt_{x_i,x_{i+1}} - \hat{t}_i}{\hat{t}_{i+1} - t^-_{i+1}}\right) && \text{(A.4)}
\end{aligned}
$$

The modification of the upper bound of the integral between Eq. A.1 and Eq. A.2 is explained by the support of y: $y \in [x + tt_{x_i,x_{i+1}}, \hat{t}_{i+1}]$, i.e., $x + tt_{x_i,x_{i+1}} \leqslant y$. Note that $x \leqslant y - tt_{x_i,x_{i+1}} \leqslant \hat{t}_{i+1} - tt_{x_i,x_{i+1}}$.

Expected value is $E(t^-_{i+1}) = \frac{\hat{t}_i + tt_{x_i,x_{i+1}} + 3 \cdot \hat{t}_{i+1}}{4}$.

A.2 Data collection campaign, data cleaning and data representativeness

Visitors from University of Lausanne (UNIL) have a unique user name for all users (unil.ch). For members of the campus, the username was associated with employee or class attribute through LDAP requests. First, 317 employees were randomly selected, 501 students were selected from 6 random classes[1] and 729 UNIL students visiting EPFL campus. For a party on campus, Vivapoly (http://vivapoly.epfl.ch/), the lists of employees and students were modified. 4493 employees were selected, corresponding to all employees who recently connected to the WiFi, and similarly all students in the selected classes, resulting in 298 IDs. After the party, the lists for students and visitors remained the same, but the list of employees returned to its original 317 IDs. The output of this process is anonymized network traces with known category of users on campus.

Data were grouped in text files on 11 different days[2], and were collected every day at a fixed time. Each time, historical data were collected individually from the CWS system of the Cisco Context Aware Mobility API with the Cisco Mobility Services Engine (MSE) Cisco (2011) that was lent to us. It was impossible to extract all data at once. This was done sequentially, MAC address by MAC address. It resulted in 2'392'973 network traces. For some users and some days, the 11 data collection campaigns would overlap over time. We cleaned the data to avoid broken daily traces related to the time of grouping in text files: if the last signal of a text file for a given individual does not appear in the next text file for the same user, we delete all signals related to this day. Similarly, for the very first signal of the first text files, we remove all signals till 3am. This way, we ensure to have complete days, and no partial sequences of signals for a day. Table A.1 shows the number of signals, IDs and days for raw data (without duplicates), and for cleaned data (without partial days).

Table A.1 – Number of network traces

	Employees	Students						Visitors
		SV-BA2	IN-BA4	GC-BA4	MA-BA2	IN-MA2	PH-BA2	UNIL
Raw data	963'294	204'516	111'199	164'203	178'339	127'911	162'954	446'344
# IDs	4140	221	125	153	168	190	176	729
# days	51	51	50	51	47	49	50	52
Clean data		203'713	110'432	162'583	177'159	127'198	161'930	
# IDs		209	114	138	152	178	158	
# days		51	50	51	47	49	50	

We applied the algorithm described in Ch. 3 to all measurements related to students with $L = 1$.

[1] Life Science, 1st year students (SV-BA2), Computer Science, 2nd year (IN-BA4), Civil Engineering, 2nd year (GC-BA4), Mathematics, 1st year (MA-BA2), Computer Science, Master students (IN-MA2), Physics, 1st year students (PH-BA2).

[2] May 23, 2012, 18:58:36; May 24, 17:10:33; May 25, 16:45:17; May 31, 10:42:24; June 8, 8:42:06; June 14, 13:06:45; June 21, 11:37:40; June 27, 10:20:43; July 2, 10:58:31; July 4, 16:31:49; July 5, 10:33:33.

The potential attractivity measure as defined in Ch. 3 is based on different data sources:

- for offices, the attractivity is based on the cumulative work percentage of the different employees of a given office, e.g., if there are 2 full time employees and one working 80% of a full time, the attractivity is defined as 2.8. No schedule is applied on offices, so this attractivity is the same all day long. These data have been provided by EPFL human resources, using SAP softwares.
- for classrooms with classes, the attractivity equals to the number of registered students to this class. The attractivity is valid only between the start and end time of the class. These data are provided by EPFL registrar's office. For the language classes, the number of registered students is not known but EPFL language center provided an estimation of 13 students per class.
- for the library and the multimedia room of the language center, the attractivity equals to the capacity, i.e., the number of seats, during opening hours. These data have been provided by EPFL library and EPFL language center.
- for the restaurants, the point-of-sale data are aggregated per quarter of hour and used as attractivity. These data have been provided by EPFL Vice Presidency for Resources and Infrastructure (Operating Support unit). Note that this is different than in Ch. 3, where the attractivity for restaurants was their capacity, i.e., their number of seats.

The attractivity and time constraints for other points of interest are not available and have to be imputed by the modeler:

- the attractivity of an office for a student must be non-zero and has been fixed to 0.1.
- the attractiviy of classrooms with classes for employees has been fixed to 2 during the class.
- the attractivity of conference rooms has been fixed to 3.
- the attractivity of the post office has been fixed to 13 during opening hours and 3 when only the ATM is available.
- the attractivity of the student union has been fixed to 3.
- the attractivity of all other points of interest has been fixed to 1, by default.

The methodology described in Ch. 3 was applied with these data and $L = 1$.

Figure A.1 shows the time spent on campus for employees based on the generated activity-episode sequences. A group of observations have a realistic duration of 8 to 9 hours. Some

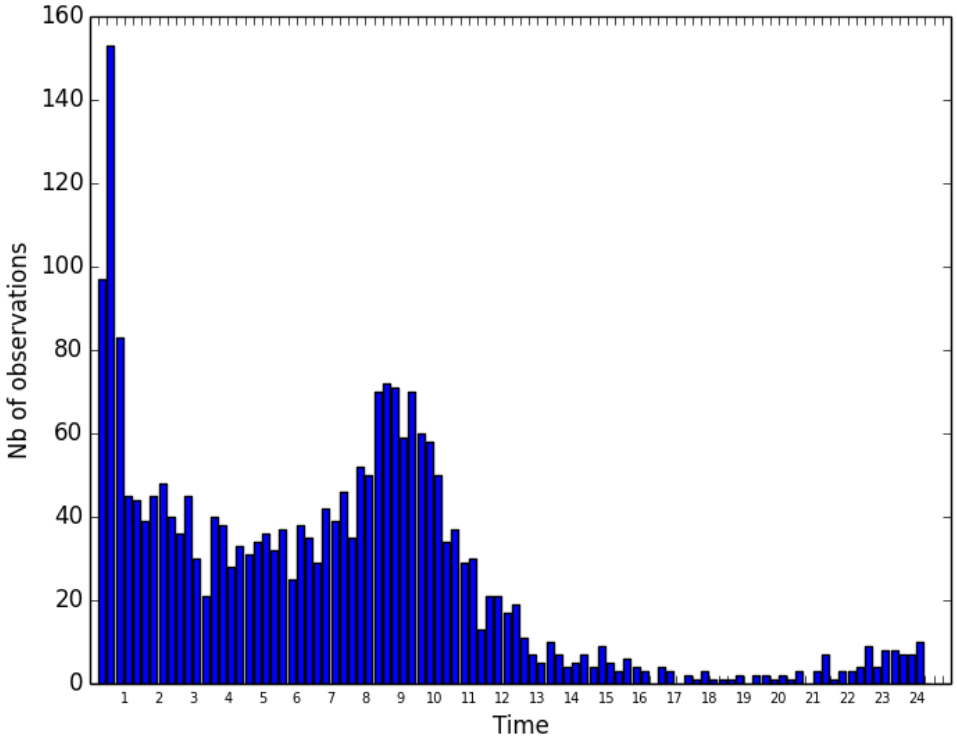


Figure A.1 – The distribution of length of activity path from activity-episode sequences generated from WiFi measurements. Y-axis represents the number of activity path with a given length.

observations have a short length, about 1 hour, or a long one, up to 24 hours. These observations are probably due to the source of data, i.e., people turning their device off in office or fix devices.

A web-based mobility survey shows that the shortest activity sequence for students and employees on campus is 2h30 per day (Tzieropoulos; 2012). In consequence, we remove all observations from WiFi traces that represent less than 2h30 on campus. Shorter observations from WiFi are assumed to be devices that are turned off part of the day, such as laptops.

Chapter 3 concludes that a density of 5.4 measurements per hour is the minimum density of measurements for generating stable and trustworthy activity-episode sequences with respect to the number of episodes, their activity type, their duration and their exact location. Consequently, we also remove activity-episode sequences with a lower mean density.

Figure A.2 shows the distribution of length of the activity-episode sequences from WiFi traces without the short sequences and without the low density observations, as described before. We see that the activity-episode sequences from WiFi traces are shorter than declared activity

sequences from the mobility survey. This might be related to an underrepresentation of short-time visitors in the web-based survey. Respondents of the mobility survey might also have declared longer activity sequences than what they really did. Cognitively incongruent answers are common for declared preferences (Bertrand and Mullainathan; 2001). Students and employees want to show a work-intensive day in their answer to the survey. It might also be that the respondents were asked to answer about a "typical day" in the recent past.

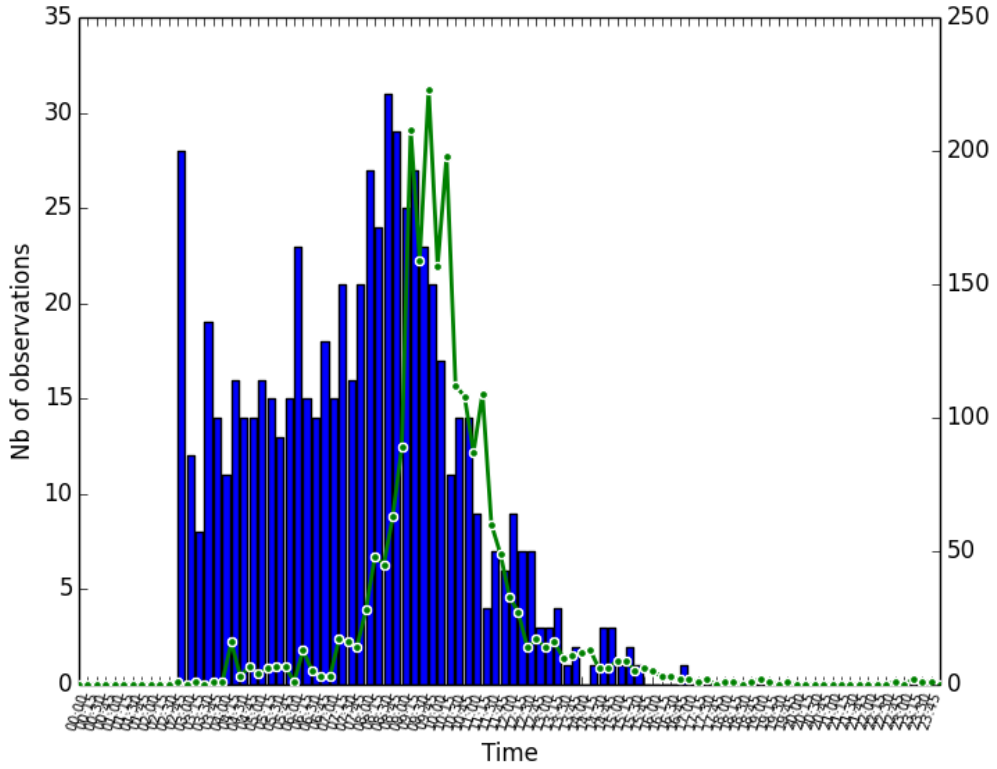


Figure A.2 – The distribution of length of activity paths from activity-episode sequences generated from WiFi measurements is represented as blue bars. The left Y-axis represents the number of activity path with a given length from WiFi measurements. The distribution of length from the mobility survey on campus (Tzieropoulos; 2012) is represented as dots and green lines. The right Y-axis represents the number of individuals with a given length from Tzieropoulos (2012). Activity sequences of less than 2h30 are removed. Activity sequences with a mean density of measurements lower than 5.4 are removed.

Figure A.3 shows when people are present on campus during the day. Blue bars describe the number of people on campus per quarter of an hour based on activity-episode sequences from WiFi traces. The green line describes the number of people on campus based on the web-based mobility survey (Tzieropoulos; 2012). People arrive earlier on campus based on the mobility survey than based on WiFi measurements.

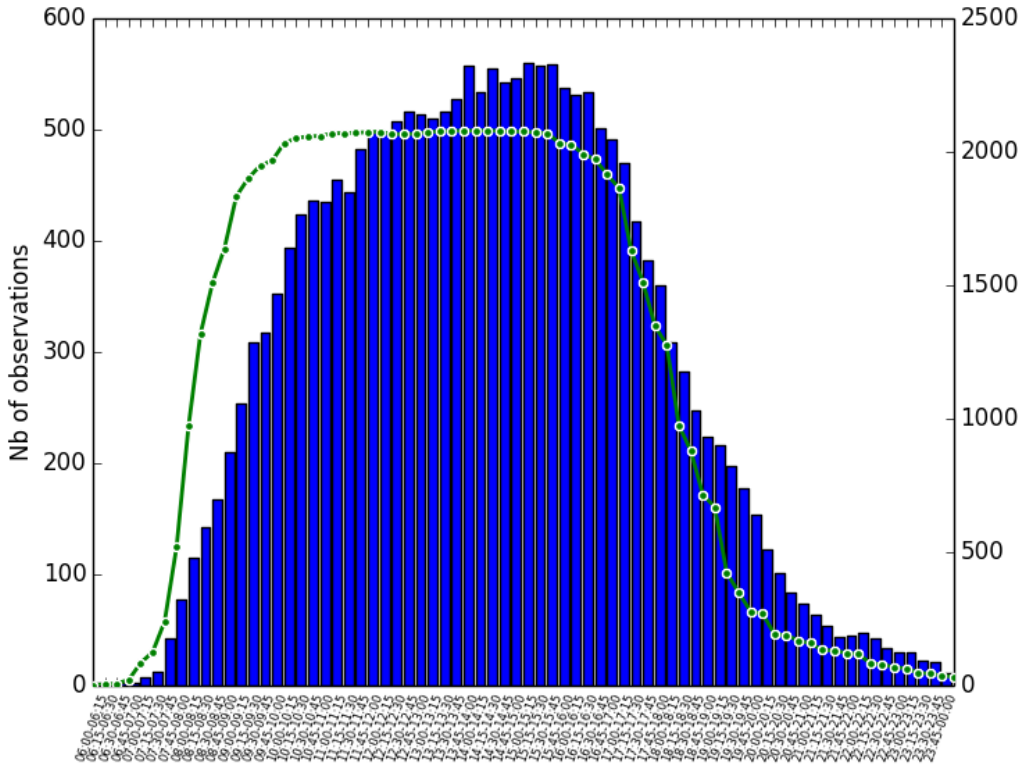


Figure A.3 – The time-of-day distribution of activity-episode sequences per quarter of an hour generated from WiFi measurements is represented as blue bars. The left Y-axis represents the number of people present at this quarter of an hour from WiFi measurements. The time-of-day distribution from (Tzieropoulos; 2012) is represented as dots and green lines. The right Y-axis represents the number of individuals at a given quarter of an hour from (Tzieropoulos; 2012).

A.3 Descriptive statistics of the WiFi traces for catering locations

As described in Danalet et al. (2014), WiFi traces have been anonymized but the category of people has been collected. Table A.2 shows the number of daily observations and the total number of individuals observed per category. Employees are overrepresented in the sample.

The number of times each catering location is chosen is described in Table A.3. The most visited catering location is L'Esplanade, very central on the campus. Le Parmentier and Le Vinci are very close and share the same kitchen; their counts of being chosen from WiFi traces are biased towards Le Parmentier, with a larger capacity and therefore a larger attractivity (see Danalet et al.; 2014). Number of visits in catering locations in the Rolex Learning Center (RLC), Le Hodler and Le Klee, are most probably underestimated due to the large attractivity of the library (see again Danalet et al.; 2014).

The walked distance to reach a catering location (Fig. A.4) is computed used a weighted

Category	Number of observations	Number of individuals
Employees	1219	145
Students, among which...	649	66
Civil engineering, Bachelor, 4th semester	131	12
Computer science, Bachelor, 4th semester	87	6
Computer science, Master, 2nd semester	53	6
Mathematics, Bachelor, 2nd semester	108	13
Life science and technology, Bachelor, 2nd semester	138	11
Physics, Bachelor, 2nd semester	132	18
Total	1868	211

Table A.2 – Number of observations and of individuals per categories of individuals.

shortest path (Danalet et al.; 2014). It takes into account the pedestrian network and the different floors on the campus. In 478 cases, distance could not be computed (previous location to the catering destination is not properly connected to the network).

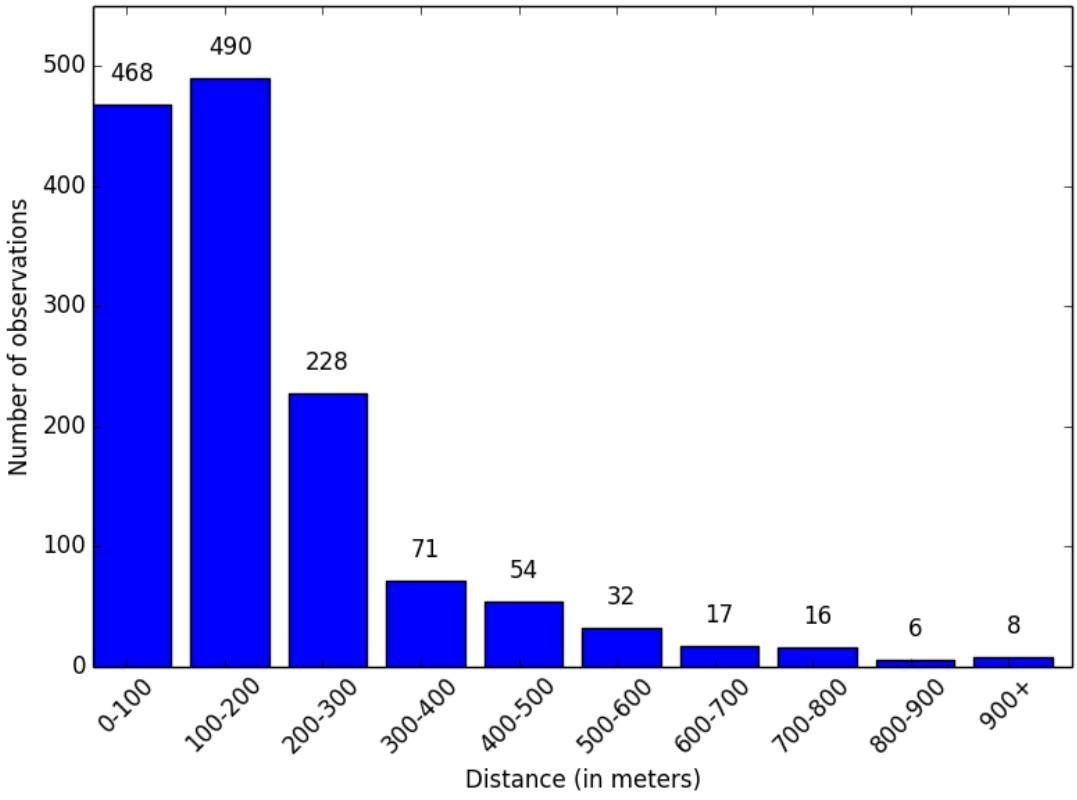


Figure A.4 – Walked distance to reach a catering location, in meters.

	Count of chosen alternatives			
Catering locations	Morning	Lunch	After lunch	*Total*
Cafeteria Cafe Le Klee	1	1	2	4
Self-service BC	46	60	40	146
Other BM	11	13	22	46
Cafeteria ELA	38	38	49	125
Cafeteria INM	3	3	7	13
Cafeteria MX	39	15	30	84
Other PH	38	7	34	79
Cafeteria L'Arcadie	19	11	8	38
Self-service L'Atlantide	73	11	51	135
Restaurant Le Copernic	0	6	0	6
Self-service Le Corbusier	17	56	0	73
Cafeteria Le Giacometti	47	44	85	176
Self-service Le Parmentier	14	68	53	135
Self-service Le Vinci	1	1	0	2
Self-service L'Esplanade	104	102	206	412
Self-service L'Ornithorynque	30	69	0	99
Caravan Pizza	18	24	22	64
Caravan Kebab	13	11	30	54
Cafeteria Satellite	37	11	87	135
Self-service Le Hodler	13	22	0	35
Restaurant Table de Vallotton	0	7	0	7
Total	562	580	728	1868

Table A.3 – Number of time each catering location is chosen in the dataset. Morning represents visits starting before 11:30, lunch visits starting between 11:30 and 14:00, and after lunch visits starting after 14:00.

More descriptive statistics about the data used in this case study are available in Tinguely (2015).

A.4 Elasticity of choice probabilities to price: detailed results

Catering locations		Static model	Dynamic model without agent effect	Dynamic model with agent effect correction	
				First choice	First choice and frequency
L'Arcadie	Employees	−1.23989	−0.985452	−3.39526	−3.57178
	Students	−2.38484	−1.83835	−4.57827	−4.75158
L'Atlantide	Employees	−1.13413	−0.895157	−3.18069	−3.32638
	Students	−2.27122	−1.77387	−4.44407	−4.625
BC	Employees	−0.936438	−0.739586	−2.61785	−2.7033
	Students	−1.57302	−1.2261	−3.08235	−3.15199
Le Copernic	Employees	−2.35353	−1.87112	−6.41268	−6.68158
	Students	−4.51551	−3.47645	−8.64327	−8.90621
Le Corbusier	Employees	−0.929999	−0.735688	−2.61505	−2.70122
	Students	−1.54962	−1.20379	−2.97714	−3.07526
ELA	Employees	−0.684047	−0.547564	−1.87449	−1.9601
	Students	−1.27976	−0.972102	−2.4586	−2.52925
L'Esplanade	Employees	−0.815134	−0.656035	−2.19422	−2.36297
	Students	−1.27379	−0.959653	−2.35676	−2.441
Le Giacometti	Employees	−0.693894	−0.553458	−1.8864	−1.96811
	Students	−1.31332	−1.00551	−2.49273	−2.58709
Le Hodler	Employees	−1.7247	−1.37671	−4.73638	−4.96936
	Students	−3.24068	−2.47713	−6.3035	−6.46166
INM	Employees	−0.763795	−0.607341	−2.08843	−2.17938
	Students	−1.45722	−1.1218	−2.79651	−2.87987
Kebab	Employees	−0.864627	−0.689636	−2.37133	−2.46379
	Students	−1.6376	−1.25415	−3.1567	−3.2466
Le Klee	Employees	−0.794574	−0.631613	−2.17519	−2.27434
	Students	−1.51289	−1.16528	−2.91007	−3.00368
MX	Employees	−0.971626	−0.767171	−2.70928	−2.8277
	Students	−1.62567	−1.26778	−3.19163	−3.31779
Ornithorynque	Employees	−0.8435	−0.664995	−2.24917	−2.37491
	Students	−1.67028	−1.30676	−3.25905	−3.3904
Le Parmentier	Employees	−0.871571	−0.695141	−2.37836	−2.49491
	Students	−1.47515	−1.12795	−2.73667	−2.8268
Pizza	Employees	−0.977648	−0.777131	−2.66259	−2.79363
	Students	−1.8736	−1.44253	−3.58497	−3.72084
Sat	Employees	−0.596256	−0.474232	−1.63987	−1.69052
	Students	−1.12653	−0.866874	−2.1905	−2.22546
Table de Vallotton	Employees	−3.9529	−3.14098	−10.5511	−10.9123
	Students	−7.58275	−5.83644	−14.3337	−14.7175
Le Vinci	Employees	−1.02426	−0.81411	−2.80212	−2.93623
	Students	−1.70861	−1.31617	−3.21097	−3.33883

Table A.4 – Average sample elasticities of choice probabilities to price

Bibliography

Aarts, H. and Dijksterhuis, A. (2000). The automatic activation of goal-directed behavioour: The case of travel habits, *Journal of Environmental Psychology* **20**(1): 75–82.
URL: *http://dx.doi.org/10.1006/jevp.1999.0156*

Abou-Zeid, M. and Ben-Akiva, M. (2012). Well-being and activity-based models, *Transportation* **39**(6): 1189–1207.
URL: *http://dx.doi.org/10.1007/s11116-012-9387-8*

Adler, T. and Ben-Akiva, M. (1979). A theoretical and empirical model of trip chaining behavior, *Transportation Research Part B* **13**(3): 243–257.
URL: *http://dx.doi.org/10.1016/0191-2615(79)90016-X*

Ahas, R., Silm, S., Järv, O., Saluveer, E. and Tiru, M. (2010). Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones, *Journal of Urban Technology* **17**(1): 3–27.
URL: *http://dx.doi.org/10.1080/10630731003597306*

Akay, A. (2012). Finite-sample comparison of alternative methods for estimating dynamic panel data models, *Journal of Applied Econometrics* **27**(7): 1189–1204.
URL: *http://dx.doi.org/10.1002/jae.1254*

Alahi, A. (2011). *Vision-Based Scene Understanding with Sparsity Promoting Priors*, PhD thesis, EPFL.
URL: *http://dx.doi.org/10.5075/epfl-thesis-5070*

Alahi, A., Bierlaire, M. and Vandergheynst, P. (2014). Robust Real-time Pedestrians Detection in Urban Environments with a Network of Low Resolution Cameras, *Transportation Research Part C* **39**: 113–128.
URL: *http://dx.doi.org/10.1016/j.trc.2013.11.019*

Anderson, I., Maitland, J., Sherwood, S., Barkhuus, L., Chalmers, M., Hall, M., Brown, B. and Muller, H. (2007). Shakra: Tracking and sharing daily activity levels with unaugmented mobile phones, *Mobile Networks and Applications* **12**(2-3): 185–199.
URL: *http://dx.doi.org/10.1007/s11036-007-0011-7*

Antonisse, R., Daly, A. and Gunn, H. (1986). The primary destination tour approach to modelling trip chains, *Proceedings of Seminar M on Transportation Planning Methods at the 14th PTRC Summer Annual Meeting*, pp. 165–177.

Arentze, T. A. and Timmermans, H. J. (2004). A learning-based transportation oriented simulation system, *Transportation Research Part B* **38**(7): 613–633.
URL: *http://dx.doi.org/10.1016/j.trb.2002.10.001*

Arikawa, M., Konomi, S. and Ohnishi, K. (2007). NAVITIME: Supporting pedestrian navigation in the real world, *IEEE Pervasive Computing* **6**(3): 21–29.
URL: *http://dx.doi.org/10.1109/MPRV.2007.61*

Arnold, S. J., Handelman, J. and Tigert, D. J. (1996). Organizational legitimacy and retail store patronage, *Journal of Business Research* **35**(3): 229–239.
URL: *http://dx.doi.org/10.1016/0148-2963(95)00128-X*

Arnold, S. J., Oum, T. H. and Tigert, D. J. (1983). Determinant Attributes in Retail Patronage: Seasonal, Temporal, Regional, and International Comparisons, *Journal of Marketing Research* **20**(2): 149.
URL: *http://dx.doi.org/10.2307/3151681*

Arnott, R., de Palma, A. and Lindsey, R. (1990). Economics of a bottleneck, *Journal of Urban Economics* **27**: 111–130.
URL: *http://dx.doi.org/10.1016/0094-1190(90)90028-L*

Arnott, R., Palma, A. D. and Lindsey, R. (1993). A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand, *The American Economic Review* **83**(1): 161–179.
URL: *http://www.jstor.org/stable/2117502*

Arulampalam, W. and Stewart, M. B. (2009). Simplified Implementation of the Heckman Estimator of the Dynamic Probit Model and a Comparison with Alternative Estimators, *Oxford Bulletin of Economics and Statistics* **71**(5): 659–681.
URL: *http://dx.doi.org/10.1111/j.1468-0084.2009.00554.x*

Aschenbruck, N., Munjal, A. and Camp, T. (2011). Trace-based mobility modeling for multi-hop wireless networks, *Computer Communications* **34**(6): 704–714.
URL: *http://dx.doi.org/10.1016/j.comcom.2010.11.002*

Atasoy, B., Glerum, A. and Bierlaire, M. (2013). Attitudes towards mode choice in Switzerland, *disP - The Planning Review* **49**(2): 101–117.
URL: *http://dx.doi.org/10.1080/02513625.2013.827518*

Axhausen, K. W., Löchl, M., Schlich, R., Buhl, T. and Widmer, P. (2007). Fatigue in long-duration travel diaries, *Transportation* **34**(2): 143–160.
URL: *http://dx.doi.org/10.1007/s11116-006-9106-4*

Axhausen, K. W., Zimmermann, A., Schönfelder, S., Rindsfüser, G. and Haupt, T. (2002). Observing the rhythms of daily life: A six-week travel diary, *Transportation* **29**(2): 95–124.
URL: *http://dx.doi.org/10.1023/A:1014247822322*

Başar, G. and Bhat, C. (2004). A parameterized consideration set model for airport choice: An application to the San Francisco Bay Area, *Transportation Research Part B* **38**: 889–904.
URL: *http://dx.doi.org/10.1016/j.trb.2004.01.001*

Balachandran, A., Voelker, G. M., Bahl, P. and Rangan, P. V. (2002). Characterizing user behavior and network performance in a public wireless LAN, *ACM SIGMETRICS Performance Evaluation Review* **30**(1): 195.
URL: *http://dx.doi.org/10.1145/511399.511359*

Balazinska, M. and Castro, P. (2003). Characterizing mobility and network usage in a corporate wireless local-area network, *ACM Mobisys*, MobiSys '03, ACM Press, New York, New York, USA, pp. 303–316.
URL: *http://dx.doi.org/10.1145/1066116.1066127*

Ball, R., Ghorpade, A., Nawarathne, K., Rui, B., Pereira, F. C., Zegras, C. and Ben-Akiva, M. (2014). Battery Patterns and Forecasting in a Large-scale Smartphone-based Travel Survey, *10th International Conference on Transport Survey Methods*, Fairmont Resort, Jamison Valley, Blue Mountains National Park, Australia.
URL: *http://dx.doi.org/10.13140/RG.2.1.3230.8641*

Balmer, M., Axhausen, K. and Nagel, K. (2006). Agent-Based Demand-Modeling Framework for Large-Scale Microsimulations, *Transportation Research Record* **1985**(1): 125–134.
URL: *http://dx.doi.org/10.3141/1985-14*

Bamberg, S., Ajzen, I. and Schmidt, P. (2003). Choice of Travel Mode in the Theory of Planned Behavior: The Roles of Past Behavior, Habit, and Reasoned Action, *Basic and Applied Social Psychology* **25**(3): 175–187.
URL: *http://dx.doi.org/10.1207/S15324834BASP2503_01*

Bauer, D., Brandle, N., Seer, S., Ray, M. and Kitazawa, K. (2009). Measurement of Pedestrian Movements: A Comparative Study on Various Existing Systems, *in* H. J. P. Timmermans (ed.), *Pedestrian Behavior: Models, Data Collection and Applications: Models, Data Collection and Applications*, Emerald Group Publishing, pp. 325–344.

Bekhor, S., Cohen, Y. and Solomon, C. (2013). Evaluating long-distance travel patterns in Israel by tracking cellular phone positions, *Journal of Advanced Transportation* **47**(4): 435–446.
URL: *http://dx.doi.org/10.1002/atr.170*

Bellomo, N., Piccoli, B. and Tosin, A. (2012). Modeling crowd dynamics from a complex system viewpoint, *Mathematical Models and Methods in Applied Sciences* **22**(2): 29 pages.
URL: *http://dx.doi.org/10.1142/S0218202512300049*

Ben-Akiva, M. and Bierlaire, M. (1999). Discrete choice methods and their applications to short-term travel decisions, *in* R. Hall (ed.), *Handbook of Transportation Science*, Operations Research and Management Science, Kluwer, Netherlands, pp. 5–34.

Ben-Akiva, M. and Bierlaire, M. (2003). Discrete choice models with applications to departure time and route choice, *in* R. Hall (ed.), *Handbook of Transportation Science, Second Edition*, Kluwer Academic Publishers, Boston/Dordrecht/London, pp. 7–37.
URL: *http://dx.doi.org/10.1007/0-306-48058-1_2*

Ben-Akiva, M. and Boccara, B. (1995). Discrete choice models with latent choice sets, *International Journal of Research in Marketing* **12**(1): 9–24.
URL: *http://dx.doi.org/10.1016/0167-8116(95)00002-J*

Ben-Akiva, M., Bowman, J. and Gopinath, D. (1996). Travel demand model system for the information era, *Transportation* **23**(3): 241–266.
URL: *http://dx.doi.org/10.1007/BF00165704*

Ben-Akiva, M. and Lerman, S. R. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press, Cambridge, MA.

Bengtsson, L., Lu, X., Thorson, A., Garfield, R. and von Schreeb, J. (2011). Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti, *PLoS Medicine* **8**(8): 1–9.
URL: *http://dx.doi.org/10.1371/journal.pmed.1001083*

Berge, G. and Peddie, S. (2010). Walking pattern in OECD/ITF countries, *in* R. Methorst (ed.), *Getting communities back on their feet, Workshop of the International Co-operation on Theories and Concepts in Traffic Safety*, The Hague, The Netherlands.

Bertrand, M. and Mullainathan, S. (2001). Do People Mean What They Say? Implications for Subjective Survey Data, *American Economic Review* **91**(2): 67–72.
URL: *http://dx.doi.org/10.1257/aer.91.2.67*

Bhargava, A. and Sargan, J. D. (1983). Estimating Dynamic Random Effects Models from Panel Data Covering Short Time Periods, *Econometrica* **51**(6): 1635.
URL: *http://dx.doi.org/10.2307/1912110*

Bhat, C. (2008). The Multiple Discrete-Continuous Extreme Value (MDCEV) Model: Role of Utility Function Parameters, Identification Considerations, and Model Extensions, *Transportation Research Part B* .
URL: *http://www.sciencedirect.com/science/article/pii/S0191261507000677*

Bhat, C. R. (2005). A multiple discrete-continuous extreme value model: formulation and application to discretionary time-use decisions, *Transportation Research Part B: Methodological* **39**(8): 679–707.
URL: *http://dx.doi.org/10.1016/j.trb.2004.08.003*

Bhat, C. R. and Koppelman, F. S. (1999). Activity-Based Modelling of Travel Demand, *Handbook of Transportation Science*, Vol. 23, Springer US, pp. 35–61.
URL: *http://dx.doi.org/10.1007/978-1-4615-5203-1_3*

Bhat, C. R. and Singh, S. K. (2000). A comprehensive daily activity-travel generation model system for workers, *Transportation Research Part A: Policy and Practice* **34**(1): 1–22.

Bierlaire, M. (2003). BIOGEME: a free package for the estimation of discrete choice models, *Proceedings of the 3rd Swiss Transportation Research Conference*, Monte Verità, Ascona, Switzerland.
URL: *http://infoscience.epfl.ch/record/117133/files/bierlaire.pdf*

Bierlaire, M., Chen, J. and Newman, J. (2013). A probabilistic map matching method for smartphone GPS data, *Transportation Research Part C* **26**: 78–98.
URL: *http://dx.doi.org/10.1016/j.trc.2012.08.001*

Bierlaire, M. and Fetiarison, M. (2009). Estimation of discrete choice models: extending BIOGEME., *Swiss Transport Research Conference (STRC)*, Monte Verità, Ascona, Switzerland.
URL: *http://www.strc.ch/conferences/2009/Bierlaire_3.pdf*

Bierlaire, M. and Frejinger, E. (2008). Route choice modeling with network-free data, *Transportation Research Part C* **16**(2): 187–198.
URL: *http://dx.doi.org/10.1016/j.trc.2007.07.007*

Bierlaire, M. and Robin, T. (2009). Pedestrians Choices, *in* H. Timmermans (ed.), *Pedestrian Behavior. Models, Data Collection and Applications*, Emerald Group Publishing Limited, pp. 1–26.

Bigano, A., Hamilton, J. M. and Tol, R. S. J. (2006). The Impact of Climate on Holiday Destination Choice, *Climatic Change* **76**(3-4): 389–406.
URL: *http://dx.doi.org/10.1007/s10584-005-9015-0*

Bolduc, D. and Ben-Akiva, M. (1991). A multinomial probit formulation for large choice sets, *Proceedings of the Sixth International Conference on Travel Behaviour*, Vol. 2, pp. 243–258.

Borgers, A. and Timmermans, H. (1986a). A Model of Pedestrian Route Choice and Demand for Retail Facilities within Inner-City Shopping Areas, *Geographical Analysis* **18**(2): 115–128.
URL: *http://dx.doi.org/10.1111/j.1538-4632.1986.tb00086.x*

Borgers, A. and Timmermans, H. (1986b). City centre entry points, store location patterns and pedestrian route choice behaviour: A microlevel simulation model, *Socio-Economic Planning Sciences* **20**(1): 25–31.
URL: *http://dx.doi.org/10.1016/0038-0121(86)90023-6*

Bowman, J. L. (1998). *The Day Activity Schedule Approach to Travel Demand Analysis*, PhD thesis, Massachusetts Institute of Technology.

Bowman, J. L. (2009). Historical Development of Activity Based Model Theory and Practice, *Traffic Engineering and Control* **50**(7): 314–318.

Bowman, J. L. and Ben-akiva, M. (1996). Activity based travel forecasting, *Conference of activity based travel forecasting (Transcript of a tutorial on activity based travel forecasting), Travel Model Improvement Program, US Department of Transportation and Environmental Protection*, New Orleans, Louisiana, pp. 1–32.

Bowman, J. L. and Ben-Akiva, M. (2001). Activity-based disaggregate travel demand model system with activity schedules, *Transportation Research Part A* **35**(1): 1–28.
URL: *http://dx.doi.org/10.1016/S0965-8564(99)00043-9*

Bradley, M., Bowman, J. L. and Griesenbeck, B. (2010). SACSIM: An applied activity-based model system with fine-level spatial and temporal resolution, *Journal of Choice Modelling* **3**(1): 5–31.
URL: *http://dx.doi.org/10.1016/S1755-5345(13)70027-7*

Büchel, D. (2004). *Développement d'une solution de navigation robuste pour l'environnement construit*, PhD thesis, EPFL.
URL: *http://infoscience.epfl.ch/record/60103*

Buchmüller, S. and Weidmann (2006). Parameters of pedestrians, pedestrian traffic and walking facilities, *Technical report*, IVT-Report Nr. 132, Institut for Transport Planning and Systems (IVT), ETHZ.

Buisson, A. (2014). *Individual activity-travel analysis based on smartphone WiFi data*, Master thesis, EPFL.
URL: *http://infoscience.epfl.ch/record/209106*

Burton, S., Tangari, A. H., Howlett, E. and Turri, A. M. (2014). How the Perceived Healthfulness of Restaurant Menu Items Influences Sodium and Calorie Misperceptions: Implications for Nutrition Disclosures in Chain Restaurants, *Journal of Consumer Affairs* **48**(1): 62–95.
URL: *http://dx.doi.org/10.1111/joca.12015*

Calabrese, F., Di Lorenzo, G., Liu, L. and Ratti, C. (2011). Estimating Origin-Destination Flows Using Mobile Phone Location Data, *IEEE Pervasive Computing* **10**(4): 36–44.
URL: *http://dx.doi.org/10.1109/MPRV.2011.41*

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J. and Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example, *Transportation Research Part C* **26**(0): 301–313.
URL: *http://dx.doi.org/10.1016/j.trc.2012.09.009*

Calabrese, F., Ferrari, L. and Blondel, V. D. (2014). Urban Sensing Using Mobile Phone Network Data: A Survey of Research, *ACM Computing Surveys* **47**(2): 1–20.
URL: *http://dx.doi.org/10.1145/2655691*

Calabrese, F., Reades, J. and Ratti, C. (2010). Eigenplaces: Segmenting Space through Digital Signatures, *IEEE Pervasive Computing* **9**(1): 78–84.
URL: *http://dx.doi.org/10.1109/MPRV.2009.62*

Cambridge Systematics Europe (1984). Estimation and Application of Disaggregate Models of Mode and Destination Choice, *Technical report*, Régie Autonome des Transports parisien, Paris.

Candamo, J., Shreve, M., Goldgof, D. B., Sapper, D. B. and Kasturi, R. (2010). Understanding transit scenes: A survey on human behavior-recognition algorithms, *IEEE Transactions on Intelligent Transportation Systems* **11**(1): 206–224.
URL: *http://dx.doi.org/10.1109/TITS.2009.2030963*

Carrel, A., Sengupta, R. and Walker, J. L. (2015). The San Francisco Travel Quality Study: Tracking Trials and Tribulations of a Transit Taker, *Technical report*, University of California, Berkeley, Berkeley, Ca.
URL: *http://www.joanwalker.com/uploads/3/6/9/5/3695513/carrel_et_al_2015_sftqs_.pdf*

Carrion, C., Pereira, F., Ball, R., Zhao, F., Kim, Y., Nawarathne, K., Zheng, N., Zegras, C. and Ben-Akiva, M. (2014). Evaluating FMS: A Preliminary Comparison with a Traditional Travel Survey, *Transporation Research Board 93rd Annual Meeting*, Washington D.C.

Carrothers, G. A. P. (1956). An Historical Bedew of the Gravity and Potential Concepts of Human Interaction, *Journal of the American Institute of Planners* **22**(2): 94–102.
URL: *http://dx.doi.org/10.1080/01944365608979229*

Cascetta, E., Nuzzolo, A., Russo, F. and Vitetta, A. (1996). A modified logit route choice model overcoming path overlapping problems: Specification and some calibration results for interurban networks, *in* J. Lesort (ed.), *Proceedings of the 13th International Symposium on Transportation and Traffic Theory*, Lyon, France, pp. 697–711.

Chen, C., Gong, H., Lawson, C. and Bialostozky, E. (2010). Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study, *Transportation Research Part A* **44**(10): 830–840.
URL: *http://dx.doi.org/10.1016/j.tra.2010.08.004*

Chen, J. (2013). *Modeling route choice behavior using smartphone data*, PhD thesis, Ecole Polytechnique Fédérale de Lausanne, Switzerland.
URL: *http://dx.doi.org/10.5075/epfl-thesis-5649*

Chen, X. and Yang, X. (2014). Does food environment influence food choices? A geographical analysis through "tweets", *Applied Geography* **51**: 82–89.
URL: *http://dx.doi.org/10.1016/j.apgeog.2014.04.003*

Chi, C. G.-Q. and Qu, H. (2008). Examining the structural relationships of destination image, tourist satisfaction and destination loyalty: An integrated approach, *Tourism Management*

29(4): 624–636.
URL: *http://dx.doi.org/10.1016/j.tourman.2007.06.007*

Childress, S. (2010). Focus Model Calibration 1.0, Denver Regional Council of Governments (DRCOG), Activity-Based Travel Model, *Technical report*, Denver Regional Council of Governments, Cambridge Systematics.

Choi, Y. K. (1999). The morphology of exploration and encounter in museum layouts, *Environment and Planning B: Planning and Design* **26**(2): 241–250.
URL: *http://dx.doi.org/10.1068/b260241*

Cisco (2011). Cisco MSE API Specification Guide - Location Service, Release 7.1, *Technical report*, Cisco.

Conti, M. and Giordano, S. (2007). Multihop Ad Hoc Networking: The Theory, *IEEE Communications Magazine* **45**(4): 78–86.
URL: *http://dx.doi.org/10.1109/MCOM.2007.343616*

Cottrill, C., Pereira, F., Zhao, F., Dias, I., Lim, H., Ben-Akiva, M. and Zegras, P. (2013). Future Mobility Survey, *Transportation Research Record: Journal of the Transportation Research Board* **2354**(07): 59–67.
URL: *http://dx.doi.org/10.3141/2354-07*

Daamen, W. (2004). *Modelling Passenger Flows in Public Transport Facilities*, PhD thesis, Delft University of Technology, The Netherlands.

Danalet, A. (2015). A Bayesian Approach to Detect Pedestrian Destination-Sequences from WiFi Signatures: Data (Transp. Res. Part C, 2014).
URL: *http://dx.doi.org/10.5281/zenodo.15798*

Danalet, A. and Bierlaire, M. (2015). Strategic sampling for activity path choice, *Technical report*, Transport and Mobility Laboratory, School of Architecture, Civil and Environmental Engineering, Ecole Polytechnique Fédŕale de Lausanne, Lausanne, Switzerland.

Danalet, A., Farooq, B. and Bierlaire, M. (2014). A Bayesian approach to detect pedestrian destination-sequences from WiFi signatures, *Transportation Research Part C* **44**: 146–170.
URL: *http://dx.doi.org/10.1016/j.trc.2014.03.015*

Danalet, A., Tinguely, L., de Lapparent, M. and Bierlaire, M. (2015). Location choice with longitudinal WiFi data, *Technical report TRANSP-OR 151110, Submitted for publication in the Journal of Choice Modeling*, Transport and Mobility Laboratory, School of Architecture, Civil and Environmental Engineering, Ecole Polytechnique Fédŕale de Lausanne, Lausanne, Switzerland.

Davidson, W., Donnelly, R., Vovsha, P., Freedman, J., Ruegg, S., Hicks, J., Castiglione, J. and Picado, R. (2007). Synthesis of first practices and operational research approaches in activity-based travel demand modeling, *Transportation Research Part A: Policy and Practice*

41(5): 464–488.
URL: *http://dx.doi.org/10.1016/j.tra.2006.09.003*

de Lapparent, M. (2009). Latent class and mixed Logit models with endogenous choice set formation based on compensatory screening rules, *in* S. Hess and A. Daly (eds), *Choice Modelling: The State-of-the-Art and the State-of-Practice*, Emerald, chapter 17.

De Vos, J., Schwanen, T., Van Acker, V. and Witlox, F. (2013). Travel and Subjective Well-Being: A Focus on Findings, Methods and Future Research Needs, *Transport Reviews* **33**(4): 421–442.
URL: *http://dx.doi.org/10.1080/01441647.2013.815665*

Deaton, A. (1985). Panel data from time series of cross-sections, *Journal of Econometrics* **30**(1-2): 109–126.
URL: *http://dx.doi.org/10.1016/0304-4076(85)90134-4*

Delafontaine, M., Versichele, M., Neutens, T. and Van de Weghe, N. (2012). Analysing spatiotemporal sequences in Bluetooth tracking data, *Applied Geography* **34**: 659–668.
URL: *http://dx.doi.org/10.1016/j.apgeog.2012.04.003*

Dellaert, B. G. C., Arentze, T. A., Bierlaire, M., Borgers, A. W. J. and Timmermans, H. J. P. (1998). Investigating Consumers' Tendency to Combine Multiple Shopping Purposes and Destinations, *Journal of Marketing Research* **35**(2): 177.
URL: *http://dx.doi.org/10.2307/3151846*

Ding, D. and Gebel, K. (2012). Built environment, physical activity, and obesity: What have we learned from reviewing the literature?
URL: *http://dx.doi.org/10.1016/j.healthplace.2011.08.021*

Diogenes, M., Greene-Roesel, R., Arnold, L. and Ragland, D. (2007). Pedestrian Counting Methods at Intersections: A Comparative Study, *Transportation Research Record* **2002**(1): 26–30.
URL: *http://dx.doi.org/10.3141/2002-04*

Doherty, S. T. and Mohammadian, A. (2011). The validity of using activity type to structure tour-based scheduling models, *Transportation* **38**(1): 45–63.
URL: *http://dx.doi.org/10.1007/s11116-010-9285-x*

Duives, D. C., Daamen, W. and Hoogendoorn, S. (2014). Trajectory Analysis of Pedestrian Crowd Movements at a Dutch Music Festival, *in* U. Weidmann, U. Kirsch and M. Schreckenberg (eds), *Pedestrian and Evacuation Dynamics 2012*, Vol. 1, Springer, pp. 151–166.
URL: *http://dx.doi.org/10.1007/978-3-319-02447-9_11*

Duives, D. C., Daamen, W. and Hoogendoorn, S. P. (2013). State-of-the-art crowd motion simulation models, *Transportation Research Part C: Emerging Technologies* **37**: 193–209.
URL: *http://dx.doi.org/10.1016/j.trc.2013.02.005*

Eash, R. (1999). Destination and Mode Choice Models for Nonmotorized Travel, *Transportation Research Record: Journal of the Transportation Research Board* **1674**: 1–8.

Edwards, D. and Griffin, T. (2013). Understanding tourists' spatial behaviour: GPS tracking as an aid to sustainable destination management, *Journal of Sustainable Tourism* **21**(4): 580–595.
URL: *http://dx.doi.org/10.1080/09669582.2013.776063*

Elwood, S., Goodchild, M. F. and Sui, D. Z. (2012). Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice, *Annals of the Association of American Geographers* **102**(3): 571–590.
URL: *http://dx.doi.org/10.1080/00045608.2011.595657*

Eriksson, L., Garvill, J. and Nordlund, A. M. (2008). Interrupting habitual car use: The importance of car habit strength and moral motivation for personal car use reduction, *Transportation Research Part F: Traffic Psychology and Behaviour* **11**(1): 10–23.
URL: *http://dx.doi.org/10.1016/j.trf.2007.05.004*

Ettema, D. (1996). *Activity based Travel Demand Modelling*, PhD thesis, Eindhoven Technical University, Holland.

Ettema, D., Ashiru, O. and Polak, J. W. (2004). Modeling timing and duration of activities and trips in response to road-pricing policies, *Transportation Research Record* **41**(1894): 1–10.

Ettema, D., Bastin, F., Polak, J. and Ashiru, O. (2007). Modelling the joint choice of activity timing and duration, *Transportation Research Part A* **41**(9): 827–841.
URL: *http://dx.doi.org/10.1016/j.tra.2007.03.001*

Ettema, D. and Timmermans, H. (2003). Modelling Departure Time Choice in the Context of Activity Scheduling Behavior, *Transportation Research Record* **1831**: 39–46.

Etter, V., Kafsi, M. and Kazemi, E. (2012). Been There , Done That: What Your Mobility Traces Reveal about Your Behavior, *Nokia Mobile Data Challenge 2012 Workshop*, June 18-19, Newcastle, UK, pp. 1–6.

Eymann, A. and Ronning, G. (1997). Microeconometric models of tourists' destination choice, *Regional Science and Urban Economics* **27**(6): 735–761.
URL: *http://dx.doi.org/10.1016/S0166-0462(97)00006-9*

Faragher, R. and Harle, R. (2015). Location Fingerprinting with Bluetooth Low Energy Beacons, *IEEE Journal on Selected Areas in Communications* **PP**(99): 1–1.
URL: *http://dx.doi.org/10.1109/JSAC.2015.2430281*

Feil, M. (2010). *Choosing the Daily Schedule: Expanding Activity-Based Travel Demand Modelling*, PhD thesis, ETH.
URL: *http://dx.doi.org/10.3929/ethz-a-006200573*

Fesenmaier, D. (1988). Integrating Activity Patterns into Destination Choice Models, *Journal of Lesiure Research* **20**(3): 175–191.

FHWA (2013). Traffic Monitoring Guide (TMG), *Technical report*, U.S. Department of Transportation, Federal Highway Administration (FHWA), Washington D.C., USA.
URL: *http://www.fhwa.dot.gov/policyinformation/tmguide/*

Flötteröd, G. and Bierlaire, M. (2013). Metropolis-Hastings sampling of paths, *Transportation Research Part B* **48**: 53–66.
URL: *http://dx.doi.org/10.1016/j.trb.2012.11.002*

Fotheringham, A. S. (1986). Modelling hierarchical destination choice, *Environment and Planning A* **18**(3): 401–418.
URL: *http://dx.doi.org/10.1068/a180401*

Fox, E. J., Montgomery, A. L. and Lodish, L. M. (2004). Consumer Shopping and Spending across Retail Formats, *The Journal of Business* **77**(S2): S25–S60.
URL: *http://dx.doi.org/10.1086/381518*

Frejinger, E. and Bierlaire, M. (2007). Capturing correlation with subnetworks in route choice models, *Transportation Research Part B* **41**(3): 363–378.
URL: *http://dx.doi.org/10.1016/j.trb.2006.06.003*

Frejinger, E. and Bierlaire, M. (2010). On Path Generation Algorithms for Route Choice Models, *in* S. H. Daly and A. (eds), *Choice Modelling: The State-of-the-Art and the State-of-Practice*, Emerald Group Publishing Limited, pp. 307–315.

Frejinger, E., Bierlaire, M. and Ben-Akiva, M. (2009). Sampling of Alternatives for Route Choice Modeling, *Transportation Research Part B* **43**(10): 984–994.
URL: *http://dx.doi.org/10.1016/j.trb.2009.03.001*

Frignani, M. Z., Auld, J., Mohammadian, A. K., Williams, C. and Nelson, P. (2010). Urban Travel Route and Activity Choice Survey (UTRACS): An Internet-Based Prompted Recall Activity Travel Survey using GPS Data, *Transportation Research Record: Journal of the Transportation Research Board* **2183**: 19–28.

Fu, X. and Lam, W. H. K. (2014). A network equilibrium approach for modelling activity-travel pattern scheduling problems in multi-modal transit networks with uncertainty, *Transportation* **41**(1): 37–55.
URL: *http://dx.doi.org/10.1007/s11116-013-9470-9*

Furuichi, M. and Koppelman, F. S. (1994). An analysis of air travelers' departure airport and destination choice behavior, *Transportation research Part A* **28**(3): 187–195.

Galland, S., Knapen, L., Yasar, A. U. H., Gaud, N., Janssens, D., Lamotte, O., Koukam, A. and Wets, G. (2014). Multi-agent simulation of individual mobility behavior in carpooling, *Transportation Research Part C: Emerging Technologies* **45**: 83–98.
URL: *http://dx.doi.org/10.1016/j.trc.2013.12.012*

Gaonkar, S., Li, J. and Choudhury, R. (2008). Micro-blog: sharing and querying content through mobile phones and social participation, ... *conference on Mobile* ... p. 174.
URL: *http://dx.doi.org/10.1145/1378600.1378620*

Gardner, B. (2009). Modelling motivation and habit in stable travel mode contexts, *Transportation Research Part F: Traffic Psychology and Behaviour* **12**(1): 68–76.
URL: *http://dx.doi.org/10.1016/j.trf.2008.08.001*

Gärling, T. and Axhausen, K. W. (2003). Introduction: Habitual travel choice, *Transportation* **30**(1): 1–11.
URL: *http://dx.doi.org/10.1023/A:1021230223001*

Goetz, M. (2012). Using Crowdsourced Indoor Geodata for the Creation of a Three-Dimensional Indoor Routing Web Application, *Future Internet* **4**(4): 575–591.
URL: *http://dx.doi.org/10.3390/fi4020575*

Goetz, M. and Zipf, A. (2011). Formal definition of a user-adaptive and length-optimal routing graph for complex indoor environments, *Geo-spatial Information Science* **14**(2): 119–128.
URL: *http://dx.doi.org/10.1007/s11806-011-0474-3*

Golob, T. F., Kitamura, R. and Long, L. (eds) (1997). *Panels for Transportation Planning*, Transportation Research, Economics and Policy, Springer US, Boston, MA.
URL: *http://dx.doi.org/10.1007/978-1-4757-2642-8*

González, M. C., Hidalgo, C. A. and Barabási, A.-L. (2008). Understanding individual human mobility patterns., *Nature* **453**(7196): 779–82.
URL: *http://dx.doi.org/10.1038/nature06958*

Gössling, S., Scott, D., Hall, C. M., Ceron, J.-P. and Dubois, G. (2012). Consumer behaviour and demand response of tourists to climate change, *Annals of Tourism Research* **39**(1): 36–58.
URL: *http://dx.doi.org/10.1016/j.annals.2011.11.002*

Greene-Roesel, R., Diogenes, M. C., Ragland, D. R. and Lindau, L. A. (2008). Effectiveness of a Commercially Available Automated Pedestrian Counting Device in Urban Environments : Comparison with Manual Counts, *TRB 2008 Annual Meeting*, Vol. c, pp. 1–16.

Grigolon, A. B., Borgers, A. W., Kemperman, A. D. and Timmermans, H. J. (2014). Vacation length choice: A dynamic mixed multinomial logit model, *Tourism Management* **41**: 158–167.
URL: *http://dx.doi.org/10.1016/j.tourman.2013.09.002*

Gupta, S. and Vovsha, P. (2013). A model for work activity schedules with synchronization for multiple-worker households, *Transportation* **40**(4): 827–845.
URL: *http://dx.doi.org/10.1007/s11116-013-9469-2*

Habib, K. M. N. (2007). *Modelling activity generation processes*, PhD thesis, University of Toronto.

Habib, K. M. N. (2011). A random utility maximization (RUM) based dynamic activity scheduling model: Application in weekend activity scheduling, *Transportation* **38**(1): 123–151.
URL: *http://dx.doi.org/10.1007/s11116-010-9294-9*

Habib, K. N., Sasic, A., Weis, C. and Axhausen, K. (2013). Investigating the nonlinear relationship between transportation system performance and daily activity-travel scheduling behaviour, *Transportation Research Part A* **49**: 342–357.
URL: *http://dx.doi.org/10.1016/j.tra.2013.01.016*

Hägerstraand, T. (1970). What about people in regional science, *Papers in Regional Science* **24**(1): 7–24.
URL: *http://dx.doi.org/10.1111/j.1435-5597.1970.tb01464.x*

Hänseler, F. S., Bierlaire, M., Farooq, B. and Mühlematter, T. (2014). A macroscopic loading model for time-varying pedestrian flows in public walking areas, *Transportation Research Part B* **69**: 60–80.
URL: *http://dx.doi.org/10.1016/j.trb.2014.08.003*

Hänseler, F. S., Bierlaire, M. and Scarpa, R. (2015). Assessing the usage and level-of-service of pedestrian facilities in train stations : A Swiss case study, *Technical report TRANSP-OR 151008*, Transport and Mobility Laboratory, School of Architecture, Civil and Environmental Engineering, Ecole Polytechnique Fédŕale de Lausanne, Lausanne, Switzerland.
URL: *http://transp-or.epfl.ch/documents/technicalReports/HaBiSc_2015.pdf*

Hansen, W. G. (1959). How Accessibility Shapes Land Use, *Journal of the American Institute of Planners* **25**(2): 73–76.
URL: *http://dx.doi.org/10.1080/01944365908978307*

Haq, S. and Luo, Y. (2012). Space Syntax in Healthcare Facilities Research: A Review, *HERD: Health Environments Research & Design Journal* **5**(4): 98–117.
URL: *http://dx.doi.org/10.1177/193758671200500409*

Hayes-Roth, B. and Hayes-Roth, F. (1979). A cognitive model of planning, *Cognitive Science* **3**(4): 275–310.
URL: *http://dx.doi.org/10.1016/S0364-0213(79)80010-5*

He, F., Wu, D., Yin, Y. and Guan, Y. (2013). Optimal deployment of public charging stations for plug-in hybrid electric vehicles, *Transportation Research Part B: Methodological* **47**: 87–101.
URL: *http://dx.doi.org/10.1016/j.trb.2012.09.007*

Heckman, J. (1978). Dummy endogenous variables in a simultaneous equation system, *Econometrica* **46**(4): 931–959.
URL: *http://dx.doi.org/10.2307/1909757*

Heckman, J. (1981). The incidental parameters problem and the problem of initial condition in estimating a discrete time-discrete data stochastic process, *in* C. Manski and D. McFad-

den (eds), *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, Cambridge, MA, pp. 179–185.

Heggie, I. and Jones, P. (1978). Defining domains for models of travel demand, *Transportation* **7**(2).
URL: *http://dx.doi.org/10.1007/BF00184635*

Helbing, D., Johansson, A. and Al-Abideen, H. Z. (2007). Dynamics of crowd disasters: An empirical study, *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* **75**(4): 1–7.
URL: *http://dx.doi.org/10.1103/PhysRevE.75.046109*

Helbing, D. and Mukerji, P. (2012). Crowd disasters as systemic failures: analysis of the Love Parade disaster, *EPJ Data Science* **1**(1): 1–40.
URL: *http://dx.doi.org/10.1140/epjds7*

Hendrich, A., Chow, M. P., Skierczynski, B. a. and Lu, Z. (2008). A 36-hospital time and motion study: how do medical-surgical nurses spend their time?, *The Permanente journal* **12**(3): 25–34.

Hess, S., Polak, J. W., Daly, A. and Hyman, G. (2007). Flexible substitution patterns in models of mode and time of day choice: New evidence from the UK and the Netherlands, *Transportation* **34**(2): 213–238.
URL: *http://dx.doi.org/10.1007/s11116-006-0011-7*

Hillier, B. (1999). *Space is the Machine: A Configurational Theory of Architecture*, Cambridge University Press, Cambridge.

Hillier, B. and Hanson, J. (1984). *The Social Logic of Space*, Cambridge University Press, Cambridge.
URL: *http://dx.doi.org/10.1017/CBO9780511597237*

Hillier, B. and Tzortzi, K. (2006). Space Syntax: The Language of Museum Space, *in* S. Macdonald (ed.), *A Companion to Museum Studies*, Blackwell Publishing Ltd, Malden, MA, USA, pp. 282–301.
URL: *http://dx.doi.org/10.1002/9780470996836.ch17*

Ho, C. and Mulley, C. (2013). Tour-based mode choice of joint household travel patterns on weekend and weekday, *Transportation* **40**(4): 789–811.
URL: *http://dx.doi.org/10.1007/s11116-013-9479-0*

Hoogendoorn, S. P. and Bovy, P. H. L. (2004). Pedestrian route-choice and activity scheduling theory and models, *Transportation Research Part B* **38**(2): 169–190.
URL: *http://dx.doi.org/10.1016/S0191-2615(03)00007-9*

Hoseini-Tabatabaei, S. A., Gluhak, A. and Tafazolli, R. (2013). A Survey on Smartphone-Based Systems for Opportunistic User Context Recognition, *ACM Computing Surveys* **45**(3): 27:1–27:51.
URL: *http://dx.doi.org/10.1145/2480741.2480744*

Hossain, A. M. and Soh, W.-S. (2015). A survey of calibration-free indoor positioning systems, *Computer Communications* **66**: 1–13.
URL: *http://dx.doi.org/10.1016/j.comcom.2015.03.001*

Hsaio, C. (2003). *Analysis of Panel Data*, Cambridge University Press, Cambridge, UK.

Hui, S., Inman, J., Huang, Y. and Suher, J. (2013). The effect of in-store travel distance on unplanned spending: applications to mobile promotion strategies, *Journal of Marketing* **77**(2): 1–16.
URL: *http://dx.doi.org/10.1509/jm.11.0436*

Hui, S. K., Bradlow, E. T. and Fader, P. S. (2009). Testing Behavioral Hypotheses Using an Integrated Model of Grocery Store Shopping Path and Purchase Behavior, *Journal of Consumer Research* **36**(3): 478–493.
URL: *http://dx.doi.org/10.1086/599046*

Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J. and Varshavsky, A. (2011). Identifying important places in people's lives from cellular network data, *in* K. Lyons, J. Hightower and E. M. Huang (eds), *Pervasive Computing*, Vol. 6696 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 133–151.
URL: *http://dx.doi.org/10.1007/978-3-642-21726-5_9*

Jenelius, E., Mattsson, L.-G. and Levinson, D. (2011). Traveler delay costs and value of time with trip chains, flexible activity scheduling and information, *Transportation Research Part B* **45**(5): 789–807.
URL: *http://dx.doi.org/10.1016/j.trb.2011.02.003*

Jiang, S., Fiore, G. A., Yang, Y., Ferreira, J., Frazzoli, E. and González, M. C. (2013). A review of urban computing for mobile phone traces, *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing - UrbComp '13*, ACM Press, New York, New York, USA.
URL: *http://dx.doi.org/10.1145/2505821.2505828*

Joh, C.-h., Arentze, T. and Timmermans, H. (2004). Activity-travel scheduling and rescheduling decision processes: empirical estimation of Aurora model, *Transportation Research Record* **1898**: 10–12.

Jones, P. M. (1979). 'HATS': a technique for investigating household decisions, *Environment and Planning A* **11**(1): 59–70.
URL: *http://dx.doi.org/10.1068/a110059*

Jones, P. M., Dix, M. C., Clarke, M. I. and Heggie, I. G. (1983). *Understanding Travel Behaviour*, Gower Publishing, Brookfield, VT USA.

Jong, G. D., Fox, J., Daly, A., Pieters, M. and Smit, R. (2004). Comparison of car ownership models, *Transport Reviews* **24**(4): 379–408.
URL: *http://dx.doi.org/10.1080/0144164032000138733*

Kalakou, S., Bierlaire, M. and Moura, F. (2014). Effects of terminal planning on passenger choices, *14th Swiss Transport Research Conference (STRC)*, Monte Verità, Ascona, Switzerland.
URL: *http://www.strc.ch/conferences/2014/Kalakou_Moura.pdf*

Kalakou, S., Psaraki-Kalouptsidi, V. and Moura, F. (2015). Future airport terminals: New technologies promise capacity gains, *Journal of Air Transport Management* **42**: 203–212.
URL: *http://dx.doi.org/10.1016/j.jairtraman.2014.10.005*

Kanda, T., Shiomi, M., Perrin, L., Nomura, T., Ishiguro, H. and Hagita, N. (2007). Analysis of people trajectories with ubiquitous sensors in a science museum, *Proceedings - IEEE International Conference on Robotics and Automation*, IEEE, Roma, pp. 4846–4853.
URL: *http://dx.doi.org/10.1109/ROBOT.2007.364226*

Kang, J. E. and Recker, W. (2013). The location selection problem for the household activity pattern problem, *Transportation Research Part B* **55**: 75–97.
URL: *http://dx.doi.org/10.1016/j.trb.2013.05.003*

Kang, J. H., Welbourne, W., Stewart, B. and Borriello, G. (2004). Extracting places from traces of locations, *Proceedings of the 2nd ACM international workshop on Wireless mobile applications and services on WLAN hotspots - WMASH '04*, WMASH '04, ACM Press, New York, New York, USA, p. 110.
URL: *http://dx.doi.org/10.1145/1024733.1024748*

Kasemsuppakorn, P. and Karimi, H. A. (2013). A pedestrian network construction algorithm based on multiple GPS traces, *Transportation Research Part C* **26**(0): 285–300.
URL: *http://dx.doi.org/10.1016/j.trc.2012.09.007*

Kazagli, E., Chen, J. and Bierlaire, M. (2014). Individual Mobility Analysis Using Smartphone Data, *in* S. Rasouli and H. Timmermans (eds), *Mobile Technologies for Activity-Travel Data Collection and Analysis*, IGI Global, chapter 12, pp. 187–208.
URL: *http://dx.doi.org/10.4018/978-1-4666-6170-7.ch012*

Kemperman, A. D., Borgers, A. W. and Timmermans, H. J. (2009). Tourist shopping behavior in a historic downtown area, *Tourism Management* **30**(2): 208–218.
URL: *http://dx.doi.org/10.1016/j.tourman.2008.06.002*

Khan, N. (2012). Analyzing patient flow: reviewing literature to understand the contribution of space syntax to improive operational efficiency in healthcare settings, *in* M. Greene, J. Reyes and A. Castro (eds), *Eight International Space Syntax Symposium*, PUC, Santiago de Chile, pp. 8183: 1–11.

Kholod, M., Nakahara, T., Azuma, H. and Yada, K. (2010). The Influence of Shopping Path Length on Purchase Behavior in Grocery Store, *in* R. Setchi, I. Jordanov, R. J. Howlett and L. C. Jain (eds), *Knowledge-Based and Intelligent Information and Engineering Systems, 14th International Conference, KES 2010, Cardiff, UK, September 8-10, 2010, Proceedings, Part III*,

Vol. 6278 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 273–280.
URL: *http://dx.doi.org/10.1007/978-3-642-15393-8_31*

Kim, T. W., Kim, Y., Cha, S. H. and Fischer, M. (2015). Automated updating of space design requirements connecting user activities and space types, *Automation in Construction* **50**: 102–110.
URL: *http://dx.doi.org/10.1016/j.autcon.2014.12.010*

Kitamura, R. (1990). Panel analysis in transportation planning: An overview, *Transportation Research Part A: General* **24**(6): 401–415.
URL: *http://dx.doi.org/10.1016/0191-2607(90)90032-2*

Kneidl, A., Hartmann, D. and Borrmann, A. (2013). A hybrid multi-scale approach for simulation of pedestrian dynamics, *Transportation Research Part C* .
URL: *http://dx.doi.org/10.1016/j.trc.2013.03.005*

Kockelman, K. M. (1998). *A Utility-Theory-Consistent System-of-Demand-Equations Approach to Household Travel Choice*, PhD thesis, University of California, Berkeley.
URL: *http://www.ce.utexas.edu/prof/kockelman/public_html/dissertation.pdf*

Kockelman, K. M. (2001). A model for time- and budget-constrained activity demand analysis.
URL: *http://dx.doi.org/10.1016/S0191-2615(99)00050-8*

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*, The MIT Press, Cambridge, MA.

Koo, S. G. M., Rosenberg, C., Chan, H.-H. and Lee, Y. C. (2003). Location discovery in enterprise-based wireless networks: case studies and applications, *Annales des Télécommunications (Annals of Telecommunications)* **58**(3-4): 531–552.
URL: *http://dx.doi.org/10.1007/BF03001027*

Krausz, B. and Bauckhage, C. (2012). Loveparade 2010: Automatic video analysis of a crowd disaster, *Computer Vision and Image Understanding* **116**(3): 307–319.
URL: *http://dx.doi.org/10.1016/j.cviu.2011.08.006*

Krueger, R., Heimerl, F., Han, Q., Kurzhals, K., Koch, S. and Ertl, T. (2015). Visual Analysis of Visitor Behavior for Indoor Event Management, *2015 48th Hawaii International Conference on System Sciences*, IEEE, pp. 1148–1157.
URL: *http://dx.doi.org/10.1109/HICSS.2015.139*

Kuutti, J., Blomqvist, K. and Sepponen, R. (2014). Evaluation of Visitor Counting Technologies and Their Energy Saving Potential through Demand-Controlled Ventilation, *Energies* **7**(3): 1685–1705.
URL: *http://dx.doi.org/10.3390/en7031685*

Lai, X. and Bierlaire, M. (2014). Specification of the cross nested logit model with sampling of alternatives for route choice models, *Technical report*, Transport and Mobility Laboratory,

Ecole Polytechnique Fédérale de Lausanne, Lausanne.
URL: *http://transp-or.epfl.ch/documents/technicalReports/LaiBier14.pdf*

Lai, X. and Bierlaire, M. (2015). Specification of the cross-nested logit model with sampling of alternatives for route choice models, *Transportation Research Part B* **80**: 220–234.
URL: *http://dx.doi.org/10.1016/j.trb.2015.07.005*

Lakshmanan, J. R. and Hansen, W. G. (1965). A retail market potential model.
URL: *http://dx.doi.org/10.1080/01944366508978155*

Lane, N. D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T. and Campbell, A. T. (2010). A survey of mobile phone sensing, *IEEE Communications Magazine* **48**(9): 140–150.
URL: *http://dx.doi.org/10.1109/MCOM.2010.5560598*

Lanir, J., Bak, P. and Kuflik, T. (2014). Visualizing Proximity-Based Spatiotemporal Behavior of Museum Visitors using Tangram Diagrams, *Computer Graphics Forum* **33**(3): 261–270.
URL: *http://dx.doi.org/10.1111/cgf.12382*

Lee, H.-Y., Yang, I.-T. and Lin, Y.-C. (2012). Laying out the occupant flows in public buildings for operating efficiency, *Building and Environment* **51**: 231–242.
URL: *http://dx.doi.org/10.1016/j.buildenv.2011.11.005*

Lemp, J. D. and Kockelman, K. M. (2012). Strategic sampling for large choice sets in estimation and application, *Transportation Research Part A: Policy and Practice* **46**(3): 602–613.
URL: *http://dx.doi.org/10.1016/j.tra.2011.11.004*

Li, S. and Lee, D. (2014). Learning Daily Activity Pattern with Probabilistic Grammar, *TRB 93rd Annual Meeting Compendium of Papers*, Washington, DC, USA, p. 17.

Liao, F., Arentze, T. and Timmermans, H. (2013). Incorporating space-time constraints and activity-travel time profiles in a multi-state supernetwork approach to individual activity-travel scheduling, *Transportation Research Part B* **55**: 41–58.
URL: *http://dx.doi.org/10.1016/j.trb.2013.05.002*

Limanond, T., Niemeier, D. and Mokhtarian, P. (2005). Specification of a tour-based neighborhood shopping model, *Transportation* **32**(2): 105–134.
URL: *http://dx.doi.org/10.1007/s11116-004-6992-1*

Liu, F., Janssens, D., Cui, J., Wang, Y., Wets, G. and Cools, M. (2014). Building a validation measure for activity-based transportation models based on mobile phone data, *Expert Systems with Applications* **41**(14): 6174–6189.
URL: *http://dx.doi.org/10.1016/j.eswa.2014.03.054 http://linkinghub.elsevier.com/retrieve/pii/S0957417414002036*

Liu, H., Darabi, H., Banerjee, P. and Liu, J. (2007). Survey of Wireless Indoor Positioning Techniques and Systems, *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)* **37**(6): 1067–1080.
URL: *http://dx.doi.org/10.1109/TSMCC.2007.905750*

Liu, X. (2013). *Activity-based pedestrian behavior simulation inside intermodal facilities*, PhD thesis, Mississippi State University.

Liu, X., Usher, J. M. and Strawderman, L. (2014). An analysis of activity scheduling behavior of airport travelers, *Computers and Industrial Engineering* **74**(1): 208–218.
URL: *http://dx.doi.org/10.1016/j.cie.2014.05.016*

Lopez-Montenegro Ramil, J., Danalet, A. and Bierlaire, M. (2013). *Visualization of pedestrian demand in a 3D graph*, Semester project, EPFL, Lausanne.

Louviere, J., Train, K., Ben-Akiva, M., Bhat, C., Brownstone, D., Cameron, T. A., Carson, R. T., Deshazo, J. R., Fiebig, D., Greene, W., Hensher, D. and Waldman, D. (2005). Recent Progress on Endogeneity in Choice Modeling, *Marketing Letters* **16**(3-4): 255–265.
URL: *http://dxd.oi.org/10.1007/s11002-005-5890-4*

Lu, Y., Peponis, J. and Zimring, C. (2009). Targeted Visibility Analysis in Buildings: Correlating Targeted Visibility Analysis with Distribution of People and Their Interactions within an Intensive Care Unit, *Seventh International Space Syntax Symposium*, Stockholm, Sweden.

Mailaender, L. (2011). Geolocation Bounds for Received Signal Strength (RSS) in Correlated Shadow Fading, *2011 IEEE Vehicular Technology Conference (VTC Fall)*, IEEE, pp. 1–6.
URL: *http://dx.doi.org/10.1109/VETECF.2011.6092847*

Manataki, I. E. and Zografos, K. G. (2009). A generic system dynamics based tool for airport terminal performance analysis, *Transportation Research Part C: Emerging Technologies* **17**(4): 428–443.
URL: *http://dx.doi.org/10.1016/j.trc.2009.02.001*

McDonald, N. (2015). Assessing the Travel of the Millennial Generation Using Pseudo-Panels, *14th International Conference on Travel Behaviour Research (IATBR)*, Windsor.

McFadden, D. (1974). The measurement of urban travel demand, *Journal of Public Economics* **3**(4): 303–328.
URL: *http://dx.doi.org/10.1016/0047-2727(74)90003-6*

McFadden, D. (1978). Modelling the choice of residential location, *in* A. Karlqvist, L. Lundqvist, F. Snickars and J. Weibull (eds), *Spatial interaction theory and residential location*, North-Holland, Amsterdam, pp. 75–96.

McFadden, D. (2001). Economic Choices, *The American Economic Review* **91**(3): 351–378.
URL: *http://www.jstor.org/stable/2677869*

McNeill, P. and Chapman, S. (2005). *Research methods*, Routledge, London.

Meneses, F. and Moreira, A. (2012). Large scale movement analysis from WiFi based location data, *2012 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, IEEE, pp. 1–9.
URL: *http://dx.doi.org/10.1109/IPIN.2012.6418885*

Miller, E. J., Roorda, M. J. and Carrasco, J. A. (2005). A tour-based model of travel mode choice, *Transportation* **32**(4): 399–422.
URL: *http://dx.doi.org/10.1007/s11116-004-7962-3*

Miller, H. J. (2010). Measuring Space-Time Accessibility Benefits within Transportation Networks: Basic Theory and Computational Procedures, *Geographical Analysis* **31**(1): 1–26.
URL: *http://dx.doi.org/10.1111/j.1538-4632.1999.tb00408.x*

Miller, H. J. (2014). Activity-Based Analysis, *in* M. M. Fischer and P. Nijkamp (eds), *Handbook of Regional Science*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 705–724.
URL: *http://dx.doi.org/10.1007/978-3-642-23430-9_106*

Millonig, A. and Maierbrugger, G. (2010). Identifying Unusual Pedestrian Movement Behaviour in Public Transport Infrastructures, *in* B. Gottfried, P. Laube, A. Klippel, N. Van de Weghe and R. Billen (eds), *MPA'10, Proceedings of the Workshop on Movement Pattern Analysis 2010*, Vol. 652, CEUR Workshop Proceedings, Zurich, Switzerland, pp. 106–110.

Millward, H., Spinney, J. and Scott, D. (2013). Active-transport walking behavior: Destinations, durations, distances, *Journal of Transport Geography* **28**: 101–110.
URL: *http://dx.doi.org/10.1016/j.jtrangeo.2012.11.012*

Mitchell, R. B. and Rapkin, C. (1954). *Urban traffic - A function of land use*, Columbia University Press, New York.

Morency, C., Trépanier Martin, M. and Demers, M. (2011). Walking to transit: An unexpected source of physical activity, *Transport Policy* **18**(6): 800–806.
URL: *http://dx.doi.org/10.1016/j.tranpol.2011.03.010*

Moussaïd, M., Helbing, D., Garnier, S., Johansson, A., Combe, M. and Theraulaz, G. (2009). Experimental study of the behavioural mechanisms underlying self-organization in human crowds., *Proceedings. Biological sciences / The Royal Society* **276**(1668): 2755–2762.
URL: *http://dx.doi.org/10.1098/rspb.2009.0405*

Mustafa, M. and Ashaari, Y. (2015). Assessing Pedestrian Behavioral Pattern at Rail Transit Terminal: State of the Art, *in* R. Hassan, M. Yusoff, A. Alisibramulisi, N. Mohd Amin and I. Zulhabri (eds), *InCIEC 2014, Proceedings of the International Civil and Infrastructure Engineering Conference 2014*, Springer Singapore, Singapore, pp. 1245–1254.
URL: *http://dx.doi.org/10.1007/978-981-287-290-6_110*

Naini, F. M., Dousse, O., Thiran, P. and Vetterli, M. (2011). Population size estimation using a few individuals as agents, *2011 IEEE International Symposium on Information Theory Proceedings*, IEEE, pp. 2499–2503.
URL: *http://dx.doi.org/10.1109/ISIT.2011.6034016*

Newell, A. and Simon, H. A. (1972). *Human Problem Solving*, Prentice-Hall, Oxford, England.

Nijland, L., Arentze, T. and Timmermans, H. (2014). Multi-day activity scheduling reactions to planned activities and future events in a dynamic model of activity-travel behavior, *Journal of Geographical Systems* **16**(1): 71–87.
URL: *http://dx.doi.org/10.1007/s10109-013-0187-2*

Nordback, K., Tufte, K., Harvey, M., McNeil, N., Stolz, E. and Liu, J. (2015). Creating a National Non-motorized Traffic Count Archive: Process and Progress, *Transportation Research Board 94th Annual Meeting*, Washington D.C., USA, p. 19.

O'Connor, A., Zerger, A. and Itami, B. (2005). Geo-temporal tracking and analysis of tourist movement, *Mathematics and Computers in Simulation* **69**(1-2): 135–150.
URL: *http://dx.doi.org/10.1016/j.matcom.2005.02.036*

Oppermann, M. (2000). Tourism Destination Loyalty, *Journal of Travel Research* **39**(1): 78–84.
URL: *http://dx.doi.org/10.1177/004728750003900110*

Ortúzar, J. D. D., Armoogum, J., Madre, J. and Potier, F. (2011). Continuous Mobility Surveys: The State of Practice, *Transport Reviews* **31**(3): 293–312.
URL: *http://dx.doi.org/10.1080/01441647.2010.510224*

Ouyang, L. Q., Lam, W. H. K., Li, Z. C. and Huang, D. (2011). Network User Equilibrium Model for Scheduling Daily Activity Travel Patterns in Congested Networks, *Transportation Research Record: Journal of the Transportation Research Board* **2254**(-1): 131–139.
URL: *http://dx.doi.org/10.3141/2254-14*

Pagliara, F. and Timmermans, H. J. P. (2009). Choice set generation in spatial contexts: a review, *Transportation Letters: The International Journal of Transportation Research* **1**(3): 181–196.
URL: *http://dx.doi.org/10.3328/TL.2009.01.03.181-196*

Palchykov, V., Mitrović, M., Jo, H.-H., Saramäki, J. and Pan, R. K. (2014). Inferring human mobility using communication patterns, *Scientific Reports* **4**: 6174.
URL: *http://dx.doi.org/10.1038/srep06174*

Papadimitriou, E., Yannis, G. and Golias, J. (2009). A critical assessment of pedestrian behaviour models, *Transportation Research Part F* **12**(3): 242–255.
URL: *http://dx.doi.org/10.1016/j.trf.2008.12.004*

Pendyala, R. M., Kitamura, R. and Reddy, D. V. G. P. (1998). Application of an activity-based travel-demand model incorporating a rule-based algorithm, *Environment and Planning B: Planning and Design* **25**(5): 753–772.
URL: *http://dx.doi.org/10.1068/b250753*

Perchoux, C., Chaix, B., Cummins, S. and Kestens, Y. (2013). Conceptualization and measurement of environmental exposure in epidemiology: Accounting for activity space related to daily mobility, *Health and Place* **21**: 86–93.
URL: *http://dx.doi.org/10.1016/j.healthplace.2013.01.005*

Philipona, C. (2002). Ne perdez pas le nord !, *Flash Informatique* **7**: 1–5.
URL: *http://flashinformatique.epfl.ch/IMG/pdf/7-2-page1.pdf*

Pinjari, A. R. and Bhat, C. (2010). A multiple discrete-continuous nested extreme value (MDCNEV) model: Formulation and application to non-worker activity time-use and timing behavior on weekdays, *Transportation Research Part B* **44**(4): 562–583.
URL: *http://dx.doi.org/10.1016/j.trb.2009.08.001*

Pinjari, A. R. and Bhat, C. R. (2011). Activity Based Travel Demand Analysis, *in* A. de Palma, R. Lindsey, E. Quinet and R. Vickerman (eds), *A Handbook of Transport Economics*, Edward Elgar Publishing Ltd, chapter 10, pp. 213–248.

Pivo, G. and Fisher, J. D. (2011). The walkability premium in commercial real estate investments, *Real Estate Economics* **39**(2): 185–219.
URL: *http://dx.doi.org/10.1111/j.1540-6229.2010.00296.x*

Popoola, O. P. and Wang, K. (2012). Video-Based Abnormal Human Behavior Recognition - A Review, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* **42**(6): 865–878.
URL: *http://dx.doi.org/10.1109/TSMCC.2011.2178594*

Prato, C. and Bekhor, S. (2006). Applying Branch-and-Bound Technique to Route Choice Set Generation, *Transportation Research Record* **1985**(1): 19–28.
URL: *http://dx.doi.org/10.3141/1985-03*

Prentow, T. S., Blunck, H., Grønbæk, K. and Kjærgaard, M. B. (2014). Estimating Common Pedestrian Routes through Indoor Path Networks using Position Traces, *2014 IEEE 15th International Conference on Mobile Data Management (MDM)*, IEEE, Brisbane, pp. 43–48.
URL: *http://dx.doi.org/10.1109/MDM.2014.11*

Quinlan, E. (2008). Conspicuous Invisibility: Shadowing as a Data Collection Strategy, *Qualitative Inquiry* **14**(8): 1480–1499.
URL: *http://dx.doi.org/10.1177/1077800408318318*

Rabe-Hesketh, S. and Skrondal, A. (2013). Avoiding biased versions of Wooldridge's simple solution to the initial conditions problem, *Economics Letters* **120**(2): 346–349.
URL: *http://dx.doi.org/10.1016/j.econlet.2013.05.009*

Rasouli, S., Shiftan, Y. and Timmermans, H. (2013). Current issues in choice modeling: Choice set specification, non-utility-maximizing behavior and discrete-continuous choice problems, *Journal of Choice Modelling* **9**(2013): 1–2.
URL: *http://dx.doi.org/10.1016/j.jocm.2014.01.001*

Rasouli, S. and Timmermans, H. (2014). Activity-based models of travel demand: promises, progress and prospects, *International Journal of Urban Sciences* **18**(1): 31–60.
URL: *http://dx.doi.org/10.1080/12265934.2013.835118*

Ratti, C. (2004). Space syntax: some inconsistencies, *Environment and Planning B: Planning and Design* **31**(4): 487–499.
URL: *http://dx.doi.org/10.1068/b3019*

Ravalet, E., Christie, D., Munafò, S. and Kaufmann, V. (2014). Analysis of walking in five Swiss cities: a quantitative and spatial approach, *14th Swiss Transport Research Conference (STRC)*, Monte Verità, Ascona, Switzerland, p. 15.
URL: *http://www.strc.ch/conferences/2014/Ravalet_EtAl_1.pdf*

Rieser-Schüssler, N. (2012). Capitalising modern data sources for observing and modelling transport behaviour, *Transportation Letters: The International Journal of Transportation Research* **4**(2): 115–128.
URL: *http://dx.doi.org/10.3328/TL.2012.04.02.115-128*

Rindfüser, G., Mühlhans, H., Doherty, S. T. and Beackmann, K. J. (2003). Tracing the planning and execution of activities and their attributes: Design and application of a hand-held scheduling process survey, *10th International Conference on Travel Behaviour Research*, August 10-14, Lucerne, Switzerland, pp. 1–31.

Robin, T. and Bierlaire, M. (2012). Modeling investor behavior, *Journal of Choice Modelling* **5**(2): 98–130.
URL: *http://dx.doi.org/10.1016/S1755-5345(13)70054-X*

Rojas, A., Branch, P. and Armitage, G. (2005). Experimental validation of the random waypoint mobility model through a real world mobility trace for large geographical areas, *Proceedings of the 8th ACM international symposium on Modeling, analysis and simulation of wireless and mobile systems - MSWiM '05*, MSWiM '05, ACM Press, New York, New York, USA, p. 174.
URL: *http://dx.doi.org/10.1145/1089444.1089474*

Roorda, M. J. (2005). *Activity-based modelling of household travel*, PhD thesis, University of Toronto.

Roorda, M. J., Miller, E. J. and Habib, K. M. (2008). Validation of TASHA: A 24-h activity scheduling microsimulation model, *Transportation Research Part A* **42**(2): 360–375.
URL: *http://dx.doi.org/10.1016/j.tra.2007.10.004*

Routledge, D. A., Repetto-Wright, R. and Howarth, C. I. (1974). A comparison of interviews and observation to obtain measures of children's exposure to risk as pedestrians., *Ergonomics* **17**(5): 623–638.
URL: *http://dx.doi.org/10.1080/00140137408931402*

Saarloos, D., Joh, C. H., Zhang, J. and Fujiwara, A. (2010). A segmentation study of pedestrian weekend activity patterns in a central business district, *Journal of Retailing and Consumer Services* **17**(2): 119–129.
URL: *http://dx.doi.org/10.1016/j.jretconser.2009.11.002*

Scarpa, R. and Thiene, M. (2005). Destination Choice Models for Rock Climbing in the North eastern Alps: A Latent-Class Approach Based on Intensity of Preferences, *Land Economics* **81**(3): 426–444.

Schadschneider, A., Klüpfel, H., Kretz, T., Rogsch, C. and Seyfried, A. (2009). Fundamentals of Pedestrian and Evacuation Dynamics, *in* A. Bazzan and F. Klügl (eds), *Multi-Agent Systems for Traffic and Transportation Engineering*, IGI Global, Hershey, Pennsylvania, USA, pp. 124–154.
URL: *http://dx.doi.org/10.4018/978-1-60566-226-8.ch006*

Schlich, R. (2004). *Verhaltenshomogene Gruppen in Längsschnitterhebungen*, PhD thesis, ETH Zurich.

Schwanen, T., Banister, D. and Anable, J. (2012). Rethinking habits and their role in behaviour change: the case of low-carbon mobility, *Journal of Transport Geography* **24**: 522–532.
URL: *http://dx.doi.org/10.1016/j.jtrangeo.2012.06.003*

Scott, D. M. and He, S. Y. (2012). Modeling constrained destination choice for shopping: a GIS-based, time-geographic approach, *Journal of Transport Geography* **23**: 60–71.
URL: *http://dx.doi.org/10.1016/j.jtrangeo.2012.03.021*

Seddighi, H. and Theocharous, A. (2002). A model of tourism destination choice: a theoretical and empirical analysis, *Tourism Management* **23**(5): 475–487.
URL: *http://dx.doi.org/10.1016/S0261-5177(02)00012-2*

Sevtsuk, A., Huang, S., Calabrese, F. and Ratti, C. (2009). *Mapping the MIT campus in real time using WiFi*, IGI Global, Hershey, PA, chapter XXII, pp. 326–338.
URL: *http://dx.doi.org/10.4018/978-1-60566-152-0.ch022*

Shephard, R. J. (2008). Is active commuting the answer to population health?, *Sports Medicine* **38**(9): 751–758.
URL: *http://dx.doi.org/10.2165/00007256-200838090-00004*

Shiftan, Y. (1998). Practical Approach to Model Trip Chaining, *Transportation Research Record: Journal of the Transportation Research Board* **1645**: 17–23.
URL: *http://dx.doi.org/10.3141/1645-03*

Shiftan, Y. (2008). The use of activity-based modeling to analyze the effect of land-use policies on travel behavior, *Annals of Regional Science* **42**(1): 79–97.
URL: *http://dx.doi.org/10.1007/s00168-007-0139-1*

Shiftan, Y. and Ben-Akiva, M. (2011). A practical policy-sensitive, activity-based, travel-demand model, *Annals of Regional Science* **47**: 517–541.
URL: *http://dx.doi.org/10.1007/s00168-010-0393-5*

Shiftan, Y., Kheifits, L. and Sorani, M. (2015). The Impact of Various Sustainable Transportation Policies on Travel and Emissions Using Activity-Based Modeling, *TRB 94th Annual Meeting Compendium of Papers*, Washigton DC, United States, p. 17.

Shobeirinejad, M., Veitch, T., Smart, J. C. R., Sipe, N. and Burke, M. (2013). Destination choice decisions of retail travellers: results from discrete choice modelling in Brisbane, *Australasian Transport Research Forum (ATRF)*, Brisbane, Queensland, Australia.

Sivakumar, A. and Bhat, C. (2007). Comprehensive, Unified Framework for Analyzing Spatial Location Choice, *Transportation Research Record: Journal of the Transportation Research Board* **2003**: 103–111.
URL: *http://dx.doi.org/10.3141/2003-13*

Small, B. K. A. (1982). The Scheduling of Consumer Activities: Work Trips, *The American Economic Review* **72**(3): 467–479.
URL: *http://www.jstor.org/stable/1831545*

Solak, S., Clarke, J. P. B. and Johnson, E. L. (2009). Airport terminal capacity planning, *Transportation Research Part B* **43**(6): 659–676.
URL: *http://dx.doi.org/10.1016/j.trb.2009.01.002*

Stewart, J. Q. (1948). Demographic Gravitation: Evidence and Applications, *Sociometry* **11**(1): 31–58.
URL: *http://www.jstor.org/stable/2785468*

Stopher, P. and Meyburg, A. (1975). *Urban transportation modeling and planning*, Lexington Books, Lexington, Mass., 1975.

Sugiyama, T., Neuhaus, M., Cole, R., Giles-Corti, B. and Owen, N. (2012). Destination and route attributes associated with adults' walking: A review.
URL: *http://dx.doi.org/10.1249/MSS.0b013e318247d286*

Swait, J. (2001). A non-compensatory choice model incorporating attribute cutoffs, *Transportation Research Part B* **35**: 903–928.
URL: *http://dx.doi.org/10.1016/S0191-2615(00)00030-8*

Tabak, V., de Vriesh, B. and Dijkstra, J. (2010). Simulation and validation of human movement in building spaces, *Environment and Planning B: Planning and Design* **37**(4): 592–609.
URL: *http://dx.doi.org/10.1068/b35127*

Tang, D. and Baker, M. (2000). Analysis of a local-area wireless network, *Proceedings of the 6th annual international conference on Mobile computing and networking*, MobiCom '00, ACM, New York, NY, USA, pp. 1–10.
URL: *http://dx.doi.org/10.1145/345910.345912*

Taylor, R. C. (2013). *Received Signal Strength-Based Localization of Non-Collaborative Emitters in the Presence of Correlated Shadowing Received Signal Strength-Based Localization of Non-Collaborative Emitters in the Presence of Correlated Shadowing*, Master Thesis, Virginia Tech.
URL: *http://hdl.handle.net/10919/23078*

Thiene, M. and Scarpa, R. (2009). Deriving and Testing Efficient Estimates of WTP Distributions in Destination Choice Models, *Environmental and Resource Economics* **44**(3): 379–395.
URL: *dx.doi.org/10.1007/s10640-009-9291-7*

Thøgersen, J. (2006). Understanding repetitive travel mode choices in a stable context: A panel study approach, *Transportation Research Part A: Policy and Practice* **40**(8): 621–638.
URL: *http://dx.doi.org/10.1016/j.tra.2005.11.004*

Timmermans, H. J. P. (1996). A stated choice model of sequential mode and destination choice behaviour for shopping trips, *Environment and Planning A* **28**(1): 173–184.
URL: *http://dx.doi.org/10.1068/a280173*

Timmermans, H., van der Hagen, X. and Borgers, A. (1992). Transportation systems, retail environments and pedestrian trip chaining behaviour: Modelling issues and applications, *Transportation Research Part B: Methodological* **26**(1): 45–59.
URL: *http://dx.doi.org/10.1016/0191-2615(92)90019-S*

Tinguely, L. (2015). *Exploiting pedestrian WiFi traces for destination choice modeling*, Master thesis, EPFL.
URL: *http://infoscience.epfl.ch/record/209732*

Tinguely, L. and Danalet, A. (2015). Destination Choice Model including panel data using WiFi localization in a pedestrian facility (dataset).
URL: *http://dx.doi.org/10.5281/zenodo.18528*

Ton, D. (2014). *NAVISTATION: a study into the route and activity location choice behaviour of departing pedestrians in train stations*, Master thesis, Delft University of Technology.

Train, K. (2003). *Discrete Choice Methods with Simulation*, Cambridge University Press, University of California, Berkeley.

Tseng, Y.-Y. and Verhoef, E. T. (2008). Value of time by time of day: A stated-preference study, *Transportation Research Part B* **42**(7-8): 607–618.
URL: *http://dx.doi.org/10.1016/j.trb.2007.12.001*

Tuduce, C. and Gross, T. (2005). A mobility model based on WLAN traces and its validation, *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, Vol. 1, pp. 664 – 674 vol. 1.
URL: *http://dx.doi.org/10.1109/INFCOM.2005.1497932*

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice, *Science* **211**(4481): 453–458.
URL: *http://dx.doi.org/10.1126/science.7455683*

Tzieropoulos, P. (2012). Mobility Observatory.
URL: *http://litep.epfl.ch/page-15350-en.html*

Ueno, J., Nakazawa, A. and Kishimoto, T. (2009). An Analysis of Pedestrian Movement in Multilevel Complex by Space Syntax Theory - In the Case of Shibuya Station, *in* D. Koch, L. Marcus and J. Steen (eds), *Proceedings of the 7th International Space Syntax Symposium*, KTH, Stockholm, pp. 118: 1–12.

Um, S. and Crompton, J. L. (1990). Attitude determinants in tourism destination choice, *Annals of Tourism Research* **17**(3): 432–448.
URL: *http://dx.doi.org/10.1016/0160-7383(90)90008-F*

van den Heuvel, J. and Hoogenraad, J. (2014). Monitoring the Performance of the Pedestrian Transfer Function of Train Stations Using Automatic Fare Collection Data, *Transportation Research Procedia* **2**: 642–650.
URL: *http://dx.doi.org/10.1016/j.trpro.2014.09.107*

van den Heuvel, J., Voskamp, A., Daamen, W. and Hoogendoorn, S. P. (2015). Using Bluetooth to Estimate the Impact of Congestion on Pedestrian Route Choice at Train Stations, *Traffic and Granular Flow '13*, Springer International Publishing, Cham, pp. 73–82.
URL: *http://dx.doi.org/10.1007/978-3-319-10629-8_9*

van der Zijpp, N. and Fiorenzo Catalano, S. (2005). Path enumeration by finding the constrained K-shortest paths, *Transportation Research Part B: Methodological* **39**(6): 545–563.
URL: *http://dx.doi.org/10.1016/j.trb.2004.07.004*

Van Nes, R., Hoogendoorn-Lanser, S. and Koppelman, F. S. (2008). Using Choice Sets for Estimation and Prediction in Route Choice, *Transportmetrica* **4**(2): 83–96.
URL: *http://dx.doi.org/10.1080/18128600808685686*

Verplanken, B., Walker, I., Davis, A. and Jurasek, M. (2008). Context change and travel mode choice: Combining the habit discontinuity and self-activation hypotheses, *Journal of Environmental Psychology* **28**(2): 121–127.
URL: *http://dx.doi.org/10.1016/j.jenvp.2007.10.005*

Versichele, M., Neutens, T., Delafontaine, M. and Van de Weghe, N. (2012). The use of Bluetooth for analysing spatiotemporal dynamics of human movement at mass events: A case study of the Ghent Festivities, *Applied Geography* **32**(2): 208–220.
URL: *http://dx.doi.org/10.1016/j.apgeog.2011.05.011*

Vickrey, W. S. (1969). Congestion Theory and Transport Investment, *The American Economic Review* **59**(2): 251–260.
URL: *http://www.jstor.org/stable/1823678*

Vovsha, P. and Bekhor, S. (1998). Link-Nested Logit Model of Route Choice: Overcoming Route Overlapping Problem, *Transportation Research Record* **1645**(1): 133–142.
URL: *http://dx.doi.org/10.3141/1645-17*

Wanalertlak, W., Lee, B., Yu, C., Kim, M., Park, S.-M. and Kim, W.-T. (2011). Behavior-based mobility prediction for seamless handoffs in mobile wireless networks, *Wireless Networks*

17(3): 645–658.
URL: *http://dx.doi.org/10.1007/s11276-010-0303-x*

Wang, D. and Li, J. (2011). A two-level multiple discrete-continuous model of time allocation to virtual and physical activities, *Transportmetrica* **7**(6): 395–416.
URL: *http://dx.doi.org/10.1080/18128602.2010.520138*

Weibull, J. W. (1976). An axiomatic approach to the measurement of accessibility, *Regional Science and Urban Economics* **6**(4): 357–379.
URL: *http://dx.doi.org/10.1016/0166-0462(76)90031-4*

Weibull, J. W. (1980). On the numerical measurement of accessibility, *Environment and Planning A* **12**(1): 53–67.
URL: *http://dx.doi.org/10.1068/a120053*

Weidmann, U., Kirsch, U. and Schreckenberg, M. E. (eds) (2014). *Pedestrian and Evacuation Dynamics 2012*, Springer.

Weiner, E. (1999). *Urban Transportation Planning in the United States: An Historical Overview*, Greenwood Publishing Group.

Weis, C. and Axhausen, K. W. (2009). Induced travel demand: Evidence from a pseudo panel data based structural equations model, *Research in Transportation Economics* **25**(1): 8–18.
URL: *http://dx.doi.org/10.1016/j.retrec.2009.08.007*

Whynes, D. K., Reedand, G. and Newbold, P. (1996). General Practitioners' Choice of Referral Destination: A Probit Analysis, *Managerial and Decision Economics* **17**(6): 587.

Woodside, A. G. and Lysonski, S. (1989). A General Model Of Traveler Destination Choice, *Journal of Travel Research* **27**(4): 8–14.
URL: *http://dx.doi.org/10.1177/004728758902700402*

Wooldridge, J. M. (2005). Simple solutions to the initial conditions problem in dynamic, non-linear panel data models with unobserved heterogeneity, *Journal of Applied Econometrics* **20**(1): 39–54.
URL: *http://dx.doi.org/10.1002/jae.770*

Wu, L. (2012). *A Tourist Behavior Model System With Multi-Faceted Dependencies and Interactions*, PhD thesis, Hiroshima University.

Wu, P. P. Y. and Mengersen, K. (2013). A review of models and model usage scenarios for an airport complex system, *Transportation Research Part A: Policy and Practice* **47**: 124–140.
URL: *http://dx.doi.org/10.1016/j.tra.2012.10.015*

Yaeli, A., Bak, P., Feigenblat, G., Nadler, S., Roitman, H., Saadoun, G., Ship, H. J., Cohen, D., Fuchs, O., Ofek-Koifman, S. and Sandbank, T. (2014). Understanding customer behavior using indoor location analysis and visualization, *IBM Journal of Research and Development*

58(5/6): 3:1–3:12.
URL: *http://dx.doi.org/10.1147/JRD.2014.2337552*

Yamamoto, T. and Kitamura, R. (1999). An analysis of time allocation to in-home and out-of-home discretionary activities across working days and non-working days, *Transportation* **26**(2): 211–230.
URL: *http://dx.doi.org/10.1023/A:1005167311075*

Yang, D. and Timmermans, H. (2015). Analysis of consumer response to fuel price fluctuations applying sample selection model to GPS panel data: Dynamics in individuals' car use, *Transportation Research Part D: Transport and Environment* **38**: 67–79.
URL: *http://dx.doi.org/10.1016/j.trd.2015.04.011*

Yang, Y., Fik, T. and Zhang, J. (2013). Modeling sequential tourist flows: Where is the next destination?, *Annals of Tourism Research* **43**: 297–320.
URL: *http://dx.doi.org/10.1016/j.annals.2013.07.005*

Yao, W., Chu, C. H. and Li, Z. (2011). Leveraging complex event processing for smart hospitals using RFID, *Journal of Network and Computer Applications* **34**(3): 799–810.
URL: *http://dx.doi.org/10.1016/j.jnca.2010.04.020*

Yoon, J., Noble, B. D. and Liu, M. (2006). Building realistic mobility models from coarse-grained traces, *in Proc. MobiSys*, ACM Press, pp. 936–5983.
URL: *http://dx.doi.org/10.1145/1134680.1134699*

Yoshimura, Y., Sobolevsky, S., Ratti, C., Girardin, F., Carrascal, J. P., Blat, J. and Sinatra, R. (2014). An analysis of visitors' behavior in The Louvre Museum: a study using Bluetooth data, *Environment and Planning B: Planning and Design* **41**(6): 1113–1131.
URL: *http://dx.doi.org/10.1068/b130047p*

Yu Quan, Deng Xiaohui and Li Ning (2011). A pedestrian speed acquisition experiment based on RFID, *Proceedings 2011 International Conference on Transportation, Mechanical, and Electrical Engineering (TMEE)*, IEEE, pp. 903–906.
URL: *http://dx.doi.org/10.1109/TMEE.2011.6199348*

Zhang, L., Zhuang, Y. and Dai, X. (2012). A Configurational study of pedestrian flows in multi-level commercial space. Case study Shanghai, *in* M. Greene, J. Reyes and A. Castro (eds), *Eighth International Space Syntax Symposium*, PUC, Santiago de Chile, pp. 8044: 1–16.

Zhu, W. and Timmermans, H. (2011). Modeling pedestrian shopping behavior using principles of bounded rationality: model comparison and validation, *Journal of Geographical Systems* **13**(2): 101–126.
URL: *http://dx.doi.org/10.1007/s10109-010-0122-8*

Zhu, W., Timmermans, H. and Wang, D. (2006). Temporal Variation in Consumer Spatial Behavior in Shopping Streets, *Journal of Urban Planning and Development* **132**(3): 166–171.
URL: *http://dx.doi.org/10.1061/(ASCE)0733-9488(2006)132:3(166)*

Zola, E. and Barcelo-Arroyo, F. (2011). A comparative analysis of the user behavior in academic WiFi networks, *Proceedings of the 6th ACM workshop on Performance monitoring and measurement of heterogeneous wireless and wired networks*, PM2HW2N '11, ACM, New York, NY, USA, pp. 59–66.
URL: *http://dx.doi.org/10.1145/2069087.2069096*

Antonin Danalet
EPFL ENAC TRANSP-OR
Station 18
CH-1015 Lausanne
Switzerland

Website: http://transp-or.epfl.ch/
http://people.epfl.ch/antonin.danalet
Phone: +41 21 6932532
E-mail: antonin.danalet@epfl.ch

Education

2011–2015	**PhD** EPFL
2009	**Master Thesis** University of Waikato, New Zealand Title: An Empirical Investigation of the Determinants of Attention to Attributes in Choice Experiments. Thesis supervisor: Prof. Riccardo Scarpa & Michel Bierlaire. *Grant of a Scholarship by the Zeno Karl Schindler Foundation.*
2003–2009	**BSc & MSc in Mathematical Sciences** EPFL

Research projects

2012–2015	**Pedestrian dynamics: flows and behavior** Swiss National Science Foundation (SNSF) grant *This project aims at developing mathematical models of pedestrian dynamics, both at aggregate and disaggregate levels.*
2011–2012	**Léman 2030: Flux piétons Gare de Lausanne** Swiss Federal Railways SBB-CFF-FFS *In the framework of a 9-month collaboration with CFF, pedestrian flows in Lausanne train station are estimated and modeled.*
2011–2012	**SIEU: Urban Energy Information System** CREM, The Ark Energy *In order to measure the effects of policies by local authorities, consolidated data about energy at the local scale is necessary, in particular statistics about mobility.*
2010–2011	**SURPRICE: Sustainable mobility through road user charging** KTH, ETHZ and EPFL *In particular on the IIVar project: Equity effects of congestion charges and intra-individual variation in preferences.*
2009–2012	**OPTIMA: Inferring Transport Mode Preferences From Attitudes** PostBus *A research project on combined mobility and factors influencing travelers in their choice of transport. Target group-specific new transport offerings and services are analyzed.*

Publications

Papers in international journal

A. Danalet, B. Farooq and M. Bierlaire. A Bayesian Approach to Detect Pedestrian Destination-Sequences from WiFi Signatures, in Transportation Research Part C: Emerging Technologies, vol. 44, p. 146 - 170, 2014. doi:10.1016/j.trc.2014.03.015

Book chapter

A. Danalet, M. Bierlaire and B. Farooq. Estimating Pedestrian Destinations Using Traces from WiFi Infrastructures, in Pedestrian and Evacuation Dynamics 2012, p. 1341-1352, 2014. doi:10.1007/978-3-319-02447-9_111

Papers in conference proceedings

L. Tinguely, A. Danalet, M. de Lapparent and M. Bierlaire. Destination Choice Model including a panel effect using WiFi localization in a pedestrian facility. 15th Swiss Transport Research Conference, Monte Verità, Ascona, Switzerland, 2015.

A. Danalet and M. Bierlaire. Importance sampling for activity path choice. 15th Swiss Transport Research Conference, Monte Verità, Ascona, Switzerland, 2015.

A. Danalet and M. Bierlaire. A path choice approach to activity modeling with a pedestrian case study. 14th Swiss Transport Research Conference, Monte Verità, Ascona, Switzerland, 2014.

A. Danalet, B. Farooq and M. Bierlaire. Towards an activity-based model for pedestrian facilities. 13th Swiss Transport Research Conference, Monte Verità, Ascona, Switerland, 2013.

S. Sahaleh, M. Bierlaire, B. Farooq, A. Danalet and F. Hänseler. Scenario Analysis of Pedestrian Flow in Public Spaces. 12th Swiss Transport Research Conference, Monte Verità, Ascona, Switzerland, 2012.

Technical reports (selection)

M. Schuler, P. Faure, S. Munafò, A. Danalet and P. Dessemontet. Amélioration de la qualité de service et évolution de la fréquentation de CarPostal, 2012 (also translated in German).

A. Danalet and S. Sahaleh. Projet de recherche sur la mobilité combinée : Rapport de l'enquête de préférences déclarées, 2012.

M. Bierlaire, A. Curchod, A. Danalet, E. Doyen and P. Faure et al. Projet de recherche sur la mobilité combinée, Rapport définitif de l'enquête de préférences révélées, 2011.

Posters (selection)

A. Danalet, M. Bierlaire and B. Farooq. Estimating Pedestrian Destinations using Traces from WiFi Infrastructures. 6th International Conference on Pedestrian and Evacuation Dynamics, Zurich, Switzerland, 2012.

A. Danalet and P. Faure. Optima: When flexible transport is able to meet a dispersed demand throughout the whole territory. Marché de la recherche regiosuisse & Colloque sur le développement régional 2011, EPFL, 2011.

Seminars (selection)

A. Danalet and M. Bierlaire. The activity path approach to activity pattern modeling. 14th International Conference on Travel Behaviour Research (IATBR), Beaumont Estate, Windsor, United Kingdom, 2015.

A. Danalet and M. Bierlaire. Activity choice modeling for pedestrian facilities: Validation on synthetic data. 3nd Symposium of the European Association for Research in Transportation (hEART) 2014, Institute for Transport Studies, University of Leeds, Leeds, United Kingdom, 2014.

M. Bierlaire, A. Danalet, F. Hänseler and M. Nikolic. Recent trends in pedestrian modeling at EPFL. Congreso Chileno de Ingeniería de Transporte, Instituto Sistemas Complejos de Ingeniería, Santiago, Chile, 2013.

A. Danalet. A path choice approach to activity modeling with a pedestrian case study. Workshop "Statistique, transport et activités", Laboratoire Jean Kuntzmann (équipe Mistis) & GAEL (Laboratoire d'Economie Appliquée de Grenoble), Grenoble, France, 2013.

A. Danalet, M. Bierlaire and B. Farooq. A Pedestrian Destination-Chain Choice Model from Bayesian Estimation of Pedestrian Activities using Sensors Data. 2nd Symposium of the European Association for Research in Transportation (hEART 2013), Stockholm, Sweden, 2013.

A. Danalet, B. Farooq and M. Bierlaire. A Bayesian Approach to Detect Pedestrian Destination-Sequences from WiFi Signatures. Eighth Triennial Symposium on Transportation Analysis (TRISTAN VIII), San Pedro de Atacama, Chile, 2013.

A. Danalet. Detecting pedestrian destinations from ubiquitous digital footprint. FCL-Talk, Future Cities Laboratory (FCL), ValueLab Asia, CREATE Tower, Singapore, 2013.

M. Thémans and A. Danalet. Interactions complexes entre infrastructures et piétons: tracking et modélisation comportementale des flux piétonniers dans les secteurs des gares. Planification et construction dans les secteurs ferroviaires, Association suisse pour l'aménagement national VLP-ASPAN, Genève, 2011.

Teaching

Lecturer

- ENAC week: Le piéton vecteur d'urbanité, au coeur de l'interdisciplinarité, 2015: Les comportements piétons, approches quantitatives

  Interdisciplinary course for architecture, civil and environmental engineering students. Specifically: lecture on quantitative data and mathematical models.

- Global issues: Mobility, 2014, 2015: Le péage urbain

  Interdisciplinary course (urban sociology and mathematical modeling), Bachelor 1st year . Specifically: lecture on congestion charging and activity based modeling.

Teaching assistant for courses

- Discrete Choice Analysis: Predicting Demand and Market Shares, 2012, 2013, 2014, 2015

  Course designed by Prof. Moshe Ben-Akiva, offered at the Massachusetts Institute of Technology (MIT). Organized in Europe by Prof. Michel Bierlaire at EPFL, with Prof. Joan Walker (University of California, Berkeley) and Prof. Daniel McFadden (University of California, Berkeley. Nobel Prize Laureate, 2000).

- Optimization and simulation (Doctoral program in Civil and Environmental Engineering), 2013.

- Mathematical modeling of behavior (Mathematics, Master in Financial Engineering), 2011, 2012.

Supervisor for Master theses

- Exploiting pedestrian WiFi traces for destination choice modeling, Master thesis, Loïc Tinguely (civil engineering), 2015
- Individual activity-travel analysis based on smartphone WiFi data, Amélie Buisson (civil engineering), 2014.
- Pedestrian flow simulation and optimization in transportation hubs (case study), Sohrab Sahaleh (civil engineering), 2011.

Supervisor for semester projects

- A destination choice model for EPFL Campus, Loïc Tinguely (civil engineering), 2014.
- Identify User's Locations of Interest from Smartphone WiFi Data, Amélie Buisson (civil engineering), 2013.
- Activities in Paléo Music Festival from Bluetooth, Elisaveta Kondratieva (communication systems), 2013.
- Visualization of pedestrian demand in a 3D graph, Javier Lopez-Montenegro Ramil (computer science), 2013.
- Dynamic estimation of pedestrian origin-destination within train stations: Exploitation of pedestrian tracking data and comparison to travel surveys, Maëlle Zimmermann (mathematics), 2012.
- Tracking Pedestrians with WiFi Traces, Yusen Bian (mathematics), 2012.
- Analysis of bus frequency in Switzerland, Suzy Polka (communication systems), 2011.

Reviewing

- Computers, Environment and Urban Systems
- IET Intelligent Transport Systems
- Transportation Research Record (TRR)
- IEEE Intelligent Transportation Systems Transactions and Magazine
- Journal of Choice Modelling

Popular science

- L'EPFL trace l'itinéraire des festivaliers du Paléo, 24 Heures, July 29, 2013 (in French)
- L'EPFL a scruté les mouvements des festivaliers au Paléo, Les News de Rouge FM et Yes FM, July 22, 2013 (in French)

- Quand le WiFi se met au service du réseau piétonnier, Flash informatique, No. 2, March 19, 2013 (in French)
- Améliorer le déplacement des usagers dans les gares, CQFD, La 1ère, Thursday November 15, 2012 (in French)
- Eviter la cohue aux heures de pointe, Touring, newspaper of the Touring Club Suisse, No. 16, September 27, 2012 (in French)
- Une revanche des piétons grâce aux outils de simulation ?, Les temps modernes, La 1ère, Monday June 11, 2012 (in French)
- CarPostal, un trait d'union entre la ville et la campagne ?, Prise de Terre, La 1ère, Saturday August 27, 2011 (in French)
- Science Q&A, weekly poll about science on EPFL homepage, about value of time, Monday April 18, 2011

Data visualization

- Chemins pour Satellite, Atlas #1, Anie Gold, 2015
- Rolex Learning Center (RLC) Pedestrian Map, Fragments de l'inachevé, Le lieu unique, Nantes, October 11-November 9, 2014
- Rolex Learning Center (RLC) Pedestrian Map, dessin embryon, de l'inachevé, halles CFF à la gare, Lausanne, May 4-19, 2013

Awards and distinctions

- Second Prize in the Image category of the 2015 ACCES Visualization Contest

  55 contributions were submitted within the School of Engineering (STI) and the School of Architecture, Civil & Environmental Engineering (ENAC) of EPFL, 35 in the Image category and 20 in the Animation category.

- ENAC Research Day 2013 Doctoral poster award - 3rd prize

  30 projects were selected based on abstracts. Accepted projects were exhibited at the Research Day and participated in the poster competition.

- Invited participation in Global Young Scientist Summit, shortlisted for the Singapore Challenge

  Invited to the 2013 Global Young Scientist Summit (GYSS) in Singapore. Shortlisted for the Singapore Challenge 2013 "Innovation for Future Cities" about envisioning opportunities for sustainable development in cities with dense populations and heavy demands on infrastructure, with a white paper on Walkable Future Cities.

Research data

L. Tinguely and A. Danalet. Destination Choice Model including panel data using WiFi localization in a pedestrian facility. Zenodo, 2015. doi:10.5281/zenodo.18528

A. Danalet. A Bayesian Approach to Detect Pedestrian Destination-Sequences from WiFi Signatures: Data (Transp. Res. Part C, 2014). Zenodo, 2015. doi:10.5281/zenodo.15798